Macroeconomic Forecasting and Variable Selection with a Very Large Number of Predictors: A Penalized Regression Approach

Yoshimasa Uematsu[1]

The Institute of Statistical Mathematics

Shinya Tanaka[2]

Otaru University of Commerce

March 3, 2017

This paper studies macroeconomic forecasting and variable selection using a folded-concave penalized regression with a very large number of predictors. The penalized regression approach leads to sparse estimates of the regression coefficients, and is applicable even if the dimensionality of the model is much larger than the sample size. The first half of the paper discusses the theoretical aspects of a folded-concave penalized regression when the model exhibits time series dependence. Specifically, we show the oracle inequality and the oracle property for ultrahigh-dimensional time-dependent regressors. The latter half of the paper shows the validity of the penalized regression using two motivating empirical applications. The first forecasts U.S. GDP with the FRED-MD data using the MIDAS regression framework, where there are more than 1000 covariates, while the sample size is at most 200. The second examines how well the penalized regression screens the hidden portfolio with around 40 stocks from more than 1800 potential stocks using NYSE stock price data. Both applications reveal that the penalized regression provides remarkable results in terms of forecasting performance and variable selection.

[1]Corresponding address: Yoshimasa Uematsu, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan. E-mail: uematsu@ism.ac.jp.

[2]Otaru University of Commerce, Department of Economics, 3-5-21 Midori, Otaru, Hokkaido 047-8501, Japan. E-mail: stanaka@res.otaru-uc.ac.jp.

Keywords: *Macroeconomic forecasting, Folded-concave penalty, Ultrahigh-dimensional time series, Mixed data sampling (MIDAS), Portfolio selection.*

JEL classification: C13, C32, C52, C53, C55

1 Introduction

Recent advancements in macroeconomic data collection have led to an increased focus on high-dimensional time series analysis. A more efficient and precise analysis can thus be realized if we elicit information appropriately from a large number of explanatory variables. However, a higher-dimensional model does not necessarily yield better performance in terms of forecasting and parameter estimation; in fact, the performance varies depending on the dimensionality and which estimation method is considered. Without appropriate dimension reduction, performance may be poor owing to accumulated estimation losses from redundant or unimportant variables. After seminal papers on factor-based (diffusion index) forecasting, such as Stock and Watson (2002), this is now common tool for forecasting with large datasets. Specifically, Stock and Watson (2012) showed that factor-based forecasts have a good performance in comparison with existing forecasting methods, including autoregressive forecast, pretest methods, Bayesian model averaging, empirical Bayes, and bagging. They concluded that it seemed difficult to outperform a factor-based forecast without introducing nonlinearity and/or time-varying parameters to a forecast model.

In this paper, we tackle the high-dimensional forecasting and estimation problem from another theoretical and empirical points of view. We employ *sparse* modeling, which can allow for *ultrahigh* dimensionality, where the number of regressors diverges sub-exponentially. The unknown sparsity can be recovered using a *folded-concave penalized regression* to pursue both prediction efficiency and variable selection consistency. In particular, we consider penalties including the smoothly clipped absolute deviation (SCAD) penalty introduced by Fan and Li (2001), the minimax concave penalty (MCP) proposed by Zhang (2010) as well

as the ℓ_1-penalty (Lasso) proposed by Tibshirani (1996). Previous studies on macroeconomic forecasting using sparse modeling include Bai and Ng (2008), De Mol et al. (2008), Kock and Callot (2015), Marsilli (2014), and Nicholson et al. (2015), but basically their estimation strategies are limited to the ℓ_1-penalty. Although the ℓ_1-penalty is expected to perform well as do the SCAD and MCP theoretically as we see in a later section, this is often insufficient in terms of model selection consistency while the SCAD and MCP can have this desirable property. Moreover, it is difficult to find a statistical theory of penalized regression estimators in time a series context.

In the first half of this paper, we provide the comprehensive theoretical properties of the penalized regression estimator under suitable conditions for macroeconometrics from the perspective of both prediction efficiency and variable selection consistency. In fact, the theoretical aspects have been explored by many recent works on statistics, including Bülmann and van de Geer (2011), Fan and Lv (2011), Fan and Lv (2013), and Loh and Wainwright (2014), as well as the references therein. However, the results of these studies are not sufficient for time series econometrics. We in this paper derive a non-asymptotic upper bound for the prediction loss called the *oracle inequality*. This ensures that the forecasting value is reliable and it is an optimal forecast in the asymptotic sense. Likewise, we also show the estimation precision of the regression coefficient and the model selection consistency, known as the *oracle property*; that is, it selects the correct subset of predictors and estimates the non-zero coefficients as efficiently as would be possible if we knew which variables were irrelevant. The oracle property provides another insight into the modeling of the variable of interest. In this regard, models can be selected by information criteria, such as the AIC and BIC. These have become popular owing to their tractability, however, they are limited when dealing with high-dimensional models because they demand an exhaustive search over all submodels. In contrast, the SCAD-type penalized regression yields simultaneous estimation and model selection, even in the ultrahigh-dimensional case.

In the second half of the paper, we shed light on the validity of the penalized regression

in macroeconometrics by introducing two empirical applications. The first one focuses on the oracle inequality. We consider to forecast quarterly U.S. real GDP with a large number of monthly predictors using MIDAS (MIxed DAta Sampling) regression framework originally proposed by Ghysels et al. (2007). Since the total number of parameters is much larger than that of observations, this situation should be treated as an ultra-high dimensional problem. In contrast to the original MIDAS model of Ghysels et al. (2007), the penalized regression enables us to forecast the quarterly GDP using a large number of monthly predictors without imposing a distributed lag structure on the regression coefficients. We find that the forecasting performance of the penalized regression is better than that of the factor-based MIDAS (F-MIDAS) regression proposed by Marcellino and Schumacher (2010) and is competitive with the nowcasting model based on the state-space representation in real-time forecasting. The second application concentrates on the oracle property. We investigate how well the penalized regression can screen a (hidden) fund manager's portfolio from large-dimensional NYSE stock price data. We construct artificial portfolios, and then we confirm the penalized regression using the SCAD-type penalty effectively detects the relevant stocks that should be contained in the portfolio. These two convincing empirical applications motivate us to apply the penalized regression to macroeconomic time series broadly.

The remainder of the paper is organized as follows. Section 2 specifies an ultrahigh-dimensional time series regression model and the estimation scheme. The statistical validity of the method is confirmed in Section 3 by deriving the oracle inequality and the oracle property. Section 4 illustrates how we can apply the penalized regression for macroeconomic time series by two empirical analyses. Section 5 concludes. The proofs and miscellaneous results are collected in the Appendix.

2 Regression Model

The regression model to be considered is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_T)^\top$ is a response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)^\top$ is a covariate matrix with $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})^\top$, $\boldsymbol{u} = (u_1, \ldots, u_T)^\top$ is an error vector, and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0A}^\top, \boldsymbol{\beta}_{0B}^\top)^\top$ is a p-dimensional unknown sparse parameter vector with $\boldsymbol{\beta}_{0A} = (\beta_{0,1}, \ldots, \beta_{0,s})^\top$ an s-dimensional vector of nonzero elements and $\boldsymbol{\beta}_{0B} = \boldsymbol{0}$. We also denote jth column vector of $\boldsymbol{X}$ by $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{Tj})^\top$. Further, we write $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)$ corresponding to the decomposition of the parameter vector. Throughout the paper, we assume that for each i, $\{x_{ti}u_t\}_t$ is a martingale difference sequence with respect to an appropriate filtration.

The objective of the paper is how we construct an efficient h-step ahead forecast value of y_{T+h} and how we select variables consistently when dimension p is much larger than T. In such cases, $\boldsymbol{X}$ may contain many irrelevant columns, so that the sparsity assumption on $\boldsymbol{\beta}_0$ may be appropriate. In this paper, we consider an *ultrahigh-dimensional* case, meaning that p diverges sub-exponentially (non-polynomially). At the same time, the degree of sparsity s may also diverge, but $s < T$ must be satisfied. The estimation procedure should select a relevant model as well as consistently estimate the parameter vector. The estimator $\hat{\boldsymbol{\beta}}$ is defined as a minimizer of the objective function

$$Q_T(\boldsymbol{\beta}) \equiv (2T)^{-1}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|p_\lambda(\boldsymbol{\beta})\|_1 \tag{2}$$

over $\boldsymbol{\beta} \in \mathbb{R}^p$, where $p_\lambda(\boldsymbol{\beta}) \equiv (p_\lambda(|\beta_1|), \ldots, p_\lambda(|\beta_p|))^\top$ and $p_\lambda(v)$, for $v \geq 0$, is a penalty function indexed by a regularization parameter $\lambda(= \lambda_T) > 0$. The penalty function p_λ takes forms such as the ℓ_1-penalty (Lasso) by Tibshirani (1996), SCAD penalty by Fan and Li (2001), and MCP by Zhang (2010). These penalties belong to a family of so-called *folded-concave penalties* because of their functional forms. The statistical properties have been developed for models with a deterministic covariate and i.i.d. Gaussian errors in the

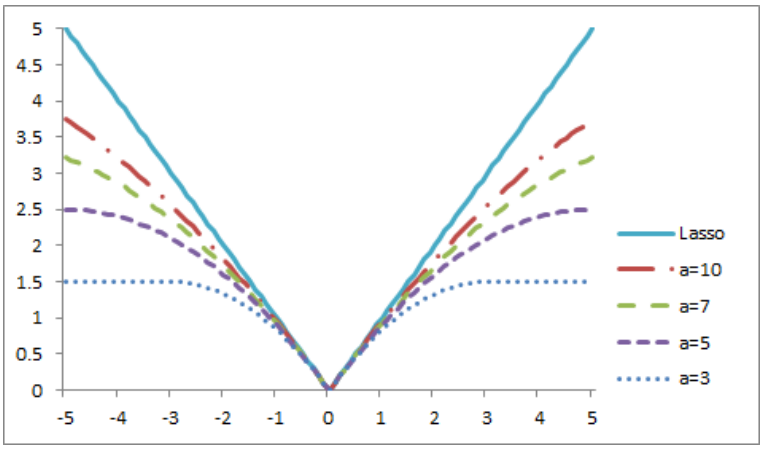

Figure 1: Shape of Folded-Concave Penalties: MCP and Lasso.

literature on high-dimensional statistics. We thoroughly investigate these properties, while relaxing the assumptions sufficiently to include many time series models.

We introduce the three penalties to be used. Let v denote a positive variable. The ℓ_1-penalty is given by $p_\lambda(v) = \lambda v$, and we then obtain $p'_\lambda(v) = \lambda$ and $p''_\lambda(v) = 0$. The SCAD penalty is defined by

$$p_\lambda(v) = \lambda v 1\{v \leq \lambda\} + \frac{a\lambda v - 0.5(v^2 + \lambda^2)}{a-1} 1\{\lambda < v \leq a\lambda\} + \frac{\lambda^2(a^2-1)}{2(a-1)} 1\{v > a\lambda\}.$$

Its derivative is

$$p'_\lambda(v) = \lambda \left\{ 1(v \leq \lambda) + \frac{(a\lambda - v)_+}{(a-1)\lambda} 1(v > \lambda) \right\},$$

for some $a > 2$. Then we have $p''_\lambda(v) = -(a-1)^{-1} 1\{v \in (\lambda, a\lambda)\}$. The MCP is defined by

$$p_\lambda(v) = \left(\lambda v - \frac{v^2}{2a} \right) 1\{v \leq a\lambda\} + \frac{1}{2} a\lambda^2 1\{v > a\lambda\}.$$

Its derivative is $p'_\lambda(v) = a^{-1}(a\lambda - v)_+$ for some $a \geq 1$. Thus, we have $p''_\lambda(v) = -a^{-1} 1\{v < a\lambda\}$. Figure 1 illustrates a shape of the MCP with several values of tuning parameters a as well as that of the Lasso.

3 Two Theoretical Results

In this section, we establish two important theoretical results, the *oracle inequality* and *oracle property* for time series models. The oracle inequality gives optimal non-asymptotic error bounds for estimation and prediction in the sense that the error bounds are of the same order of magnitude up to a logarithmic factor as those we would have if we a priori knew the relevant variables (Bülmann and van de Geer, 2011). This result strongly supports the use of penalized regressions in terms of forecasting accuracy, even in ultrahigh-dimensional spaces. Note that we should remark that the inequality provides no information on model selection consistency; that is, it is not clear whether the penalized regression correctly distinguishes the relevant variables contained in the true model from the irrelevant ones. This issue is then addressed by the oracle property, which, in turn, states that the estimator exhibits model selection consistency. The existing results have shown the oracle inequality and the oracle properly under i.i.d. Gaussian errors and deterministic covariates, but in the paper we extend these results to apply to time series models.

Assumption 1 We have $\log p = O(T^{\delta})$ and $s = O(T^{\delta_0})$ for some constants $\delta, \delta_0 \in (0, 1)$.

Assumption 2 Penalty function $p_\lambda(\cdot)$ is characterized as follows:

(a) $p_\lambda(v)$ is concave in $v \in [0, \infty)$ with $p_\lambda(0) = 0$,

(b) $p_\lambda(v)$ is nondecreasing, but $v \mapsto p_\lambda(v)/v$ $(v \neq 0)$ is nonincreasing in $v \in [0, \infty)$,

(c) $p_\lambda(v)$ has a continuous derivative $p'_\lambda(v)$ for $v \in (0, \infty)$ with $p'_\lambda(0+) = \lambda$,

(d) There exists $\mu > 0$ such that $p_\lambda(v) + \mu v^2/2$ is convex in $v \in [0, \infty)$.

Assumption 1 means that the dimensionality of the model, p, diverges sub-exponentially as T goes to infinity. Assumption 2 determines a family of folded-concave penalties that bridges ℓ_0- and ℓ_1-penalties. The SCAD and MCP are included in this family. The ℓ_1-penalty also satisfies this as the boundary of this class. It is known from Lemmas 6 and 7 of

Loh and Wainwright (2014) that (d) is true provided that $\mu \geq 1/(a-1)$ for the SCAD and $\mu \geq 1/a$ for the MCP.

We define the gradient vector and Hessian matrix of $(2T)^{-1}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$ as $G_T(\boldsymbol{\beta}) \equiv -\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})/T$ and $\boldsymbol{H}_T \equiv \boldsymbol{X}^\top\boldsymbol{X}/T$, respectively. Denoting $\boldsymbol{G}_{0T} \equiv G_T(\boldsymbol{\beta}_0)$, we may write

$$\boldsymbol{G}_{0T} = -\frac{1}{T}\begin{pmatrix}\boldsymbol{X}_A^\top\boldsymbol{u}\\ \boldsymbol{X}_B^\top\boldsymbol{u}\end{pmatrix} \equiv \begin{pmatrix}\boldsymbol{G}_{0AT}\\ \boldsymbol{G}_{0BT}\end{pmatrix}, \quad \boldsymbol{H}_T = \frac{1}{T}\begin{pmatrix}\boldsymbol{X}_A^\top\boldsymbol{X}_A & \boldsymbol{X}_A^\top\boldsymbol{X}_B\\ \boldsymbol{X}_B^\top\boldsymbol{X}_A & \boldsymbol{X}_B^\top\boldsymbol{X}_B\end{pmatrix} \equiv \begin{pmatrix}\boldsymbol{H}_{AAT} & \boldsymbol{H}_{ABT}\\ \boldsymbol{H}_{BAT} & \boldsymbol{H}_{BBT}\end{pmatrix}.$$

3.1 Oracle inequality

We derive optimal non-asymptotic error bounds for estimation and prediction called the oracle inequality. In the literature, Bülmann and van de Geer (2011, Ch. 6) presented a complete guide for the inequality using the ℓ_1-penalty with fixed predictors and i.i.d. normal errors. We extend the result in two ways. First, the inequality holds for the general model (1). Second, we prove the asymptotic equivalence of ℓ_1- and the other folded-concave penalties characterized by Assumption 2 in the sense that they satisfy the same rate. This indicates that the forecasting performance is asymptotically equivalent, irrespective of the folded-concave penalty used. We first derive the bounds under two high-level assumptions in Section 3.1.1. We next consider the conditions under which the two high-level assumptions are actually verified in a reasonable time series setting in Section 3.1.2. Related studies are introduced in Appendix A.9.

3.1.1 General result

We start with general but high-level assumptions:

Assumption 3 There are a sequence $\lambda = o(1)$ and a positive constant c_1 such that $\mathcal{E}_1^c$, the complement of the event $\mathcal{E}_1 = \{\|\boldsymbol{G}_{0T}\|_\infty \leq \lambda/2\}$, satisfies $P(\mathcal{E}_1^c) = O(p^{-c_1})$.

Assumption 4 There are a diverging sequence $m = o(T)$ and positive constants c_2 and $\gamma > \mu/2$ such that $\mathcal{E}_2^c$, the complement of the event $\mathcal{E}_2 = \left\{\min_{\boldsymbol{v}\in\mathbb{R}^p,\ \|\boldsymbol{v}\|_0\leq m} T^{-1}\|\boldsymbol{X}\boldsymbol{v}\|_2^2/\|\boldsymbol{v}\|_2^2 \geq \gamma\right\}$, satisfies $P(\mathcal{E}_2^c) = O(\exp(-c_2 T))$.

Assumption 3 requires that the gradient vector $\boldsymbol{G}_{0T}$ to behave less fluctuate and converge to zero with an appropriate rate determined by λ. For example, we should set $\lambda = O((\log p/T)^{1/2})$ for the case when $\boldsymbol{u}$ is i.i.d. normal and $\boldsymbol{X}$ is deterministic. Assumption 4 is a stochastic version of the *restricted strong convexity* studied by Negahban et al. (2012). This prevents the minimum eigenvalue of the sub-matrix of Hessian matrix $\boldsymbol{H}_T$ from being too small. These two assumptions fully control the randomness of the regression model, meaning that irrespective of the dependence structure the model possesses, Theorem 1 below holds as long as they are satisfied. The problem is what reasonable conditions on $\boldsymbol{X}$ and $\boldsymbol{u}$ satisfy Assumption 3 and 4. In fact, these can easily be verified for i.i.d. Gaussian $\boldsymbol{u}$ and deterministic $\boldsymbol{X}$. However, we may anticipate that it becomes quite unclear whether these assumptions hold or not once the model departs from such simple settings.

Under the assumptions listed above, we can derive the following result:

Theorem 1 *Let Assumptions 1–4 hold. Then, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q_T(\boldsymbol{\beta})$ on $\{\boldsymbol{\beta}\in\mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq m - s\}$ such that, with probability at least $1 - O(p^{-c_1}) - O(\exp(-c_2 T))$, the following hold:*

(a) (Estimation error in ℓ_2-norm) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq \dfrac{6s^{1/2}\lambda}{2\gamma - \mu}$,

(b) (Estimation error in ℓ_1-norm) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq \dfrac{24s\lambda}{2\gamma - \mu}$,

(c) (Prediction loss) $T^{-1/2}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2 \leq \dfrac{9s^{1/2}\lambda}{(2\gamma - \mu)^{1/2}}$.

If $2\gamma - \mu$ is assumed to be fixed, the error bounds converge to zero as long as λ goes to zero relatively faster than $s^{1/2}$ or s. In a simple setting with i.i.d. Gaussian u_t and fixed X_t, it is known that λ should be given by $O((\log p/T)^{1/2})$ as mentioned before, leading to the

explicit convergence rates $O((s \log p/T)^{1/2})$. This goes to zero provided that $\delta + \delta_0 < 1$. We observe later that the rates become slightly slower in a time series setting. Result (c) exhibits an optimal bound for the prediction loss in the ℓ_2-norm in the sense of Bickel et al. (2009). This result justifies using any penalty function specified by Assumption 2 when the aim is forecasting in the ultrahigh dimension. To understand the result, we consider a simplification in model (1) such that $\boldsymbol{X}$ is deterministic, $\boldsymbol{u}$ is i.i.d. with a unit variance, and $s = p < T$. Then, the squared risk of the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ becomes

$$T^{-1} \mathrm{E} \|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0)\|_2^2 = T^{-1} \mathrm{E}[\boldsymbol{u}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}] = T^{-1} \mathrm{tr} \boldsymbol{I} = s/T.$$

Consider the case $p \geq T > s$. If we knew the true model A, we could choose the correct s variables from $\boldsymbol{X}$, leading to the risk s/T. However, since A is unknown, the additional logarithm factor, which is regarded as the price to pay for not knowing A, is inserted.

3.1.2 When does the general result hold?

Theorem 1 has established the non-asymptotic error bounds for the penalized regression estimators and prediction error under general, yet high-level, assumptions. Specifically, Assumptions 3 and 4 should be verified for each model we attempt to employ. Here we consider a specific time series model. To consider a specific dependent model, we first strengthen the assumption on dimensionality:

Assumption 5 The dimensionality is given by $\log p = \phi T^{\delta}$ and $s = \phi_0 T^{\delta_0}$ for some positive constants ϕ, ϕ_0, δ, and δ_0 such that $\delta + \delta_0 < 1$.

In order to specify the processes of $\boldsymbol{X}$ and $\boldsymbol{u}$, we further assume in the same manner as Ahn and Horenstein (2013) that the covariate $\boldsymbol{X}$ and the error $\boldsymbol{u}$ are given by

$$\boldsymbol{X} = \boldsymbol{R}_X^{1/2} \boldsymbol{Z}_X \boldsymbol{\Sigma}_X^{1/2}, \quad \boldsymbol{u} = \sigma_u \boldsymbol{R}_u^{1/2} \boldsymbol{z}_u, \tag{3}$$

where the random matrix $\boldsymbol{Z}_X \in \mathbb{R}^{T \times p}$, random vector $\boldsymbol{z}_u \in \mathbb{R}^T$, and deterministic matrices $\boldsymbol{R}_X \in \mathbb{R}^{T \times T}$, $\boldsymbol{R}_u \in \mathbb{R}^{T \times T}$, and $\boldsymbol{\Sigma}_X \in \mathbb{R}^{p \times p}$ are characterized by the following assumption:

Assumption 6 The following conditions hold:

(a) The entries of $\boldsymbol{Z}_X$ and $\boldsymbol{z}_u$ are i.i.d. standard normal random variables.

(b) $\boldsymbol{R}_X$, $\boldsymbol{R}_u$, and $\boldsymbol{\Sigma}_X$ are symmetric and positive definite non-random matrices, the minimum eigenvalues of which are bounded from below by positive constants c_{R_X}, c_{R_u}, and c_Σ, respectively. In addition, we set $c_R = c_{R_X} \wedge c_{R_u}$ and $\sigma_u > 0$.

(c) $\boldsymbol{R}_X^{1/2} \equiv \left(r_{st}^{(X)}\right)$ and $\boldsymbol{R}_u^{1/2} \equiv \left(r_{st}^{(u)}\right)$ are lower triangular matrices whose elements satisfy $r_{tt}^{(X)} = r_{tt}^{(u)} = 1$ and $R_{sT}^{(X)} \vee R_{sT}^{(u)} = O(1)$ for all s, where $R_{sT}^{(X)} = \sum_{t=1}^{T}(r_{st}^{(X)})^2$ and $R_{sT}^{(u)} = \sum_{t=1}^{T}(r_{st}^{(u)})^2$. $\boldsymbol{\Sigma}_X^{1/2} \equiv \left(\sigma_{ij}^{(X)}\right)$ is a positive definite matrix that satisfies $\sigma_{ii}^{(X)} = 1$ and $\Sigma_{pj}^{(X)} < \infty$ for all j, where $\Sigma_{pj}^{(X)} = \sum_{i=1}^{p}(\sigma_{ij}^{(X)})^2$.

Gaussianity in condition (a) can be weakened to sub-Gaussianity. Matrices in condition (c) are defined based on the Cholesky decomposition and Spectral decomposition under condition (b). Model (3) with Assumption 6 covers a wide range of time series processes with cross-sectional dependences. A simple example of $\boldsymbol{R}_X^{1/2}$ and $\boldsymbol{\Sigma}_X^{1/2}$ is given by setting $r_{t,t-1}^{(X)} = \theta_r$ and $\sigma_{i,i-1}^{(X)} = \varphi_\sigma$ for some constants θ_r and φ_σ satisfying $|\theta_r| < 1$ and $|\varphi_\sigma| < \infty$ with other entries all zero. Obviously, this formulation satisfies condition (c) with reducing model (3) to an MA(1) process. Other weak stationary processes with cross-sectional dependences can be expressed in a similar manner.

Proposition 1 *Let Assumptions 5 and 6 hold with* $\lambda = c_0 \log(pT)(\log p/T)^{1/2}$, *with choosing positive constant* c_0 *such that* $c_0 \geq 16c_{xu}$, *where* $c_{xu} = \limsup_T \max_{t,i}\{R_{tT}^{(X)}\Sigma_{pi}, R_{tT}^{(u)}\sigma_u\} < \infty$. *Then, Assumption 3 is satisfied with* $P(\mathcal{E}_1^c) \leq 6p^{-1}$.

Proposition 2 *Let Assumptions 5 and 6 hold with* $m \leq \phi T^{1-\delta}$ *and* $\phi^2 < 1/2$. *Then, Assumption 4 is satisfied with* $\gamma = c_G c_R/9$ *and* $P(\mathcal{E}_2^c) \leq 2\exp(-c_2 T)$, *where* $c_2 = 1/2 - \phi^2$.

Combining Propositions 1 and 2 leads to the non-asymptotic bounds in the time series setting specified by Assumptions 5 and 6.

Corollary 1 *Let Assumptions 2, 5, and 6 hold with the constants being the same as in Propositions 1 and 2. Then, there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *of* $Q_T(\boldsymbol{\beta})$ *on* $\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq m - s\}$ *such that, with probability at least* $1 - 6p^{-1} - 2\exp\{-(1/2 - \phi^2)T\}$*, the the error bounds (a)–(c) of Theorem 1 hold.*

Corollary 1 does not always imply the consistency. Once the condition $\delta + \delta_0 < 1$ in Assumption 5 is strengthened to $3\delta + \delta_0 < 1$, the bounds of (a) and (c), given by $s^{1/2}\lambda = O\left(T^{\delta_0/2}\log(pT)(\log p/T)^{1/2}\right)$, converge to zero. Similarly, adding the condition $3\delta+2\delta_0 < 1$ entails the bound of (b) converges to zero.

Compared to the conventional rate, $O\left((s\log p/T)^{1/2}\right)$, obtained with i.i.d. normal errors and fixed covariates, a slightly slower rate $O\left((\log pT)(s\log p/T)^{1/2}\right)$ arises for our time series model. We can interpret the additional factor $\log(pT)$ as an extra cost of departure from the independent Gaussian world. To understand this, the point is the behavior of the process $\{x_{ti}u_t\}$ for each i. If u_t is i.i.d. Gaussian and x_{ti} is deterministic, $\{x_{ti}u_t\}$ becomes a sequence of independent normal random variables. Hence, it is easy to control the tail probability $P(\|\boldsymbol{G}_{0T}\|_\infty > \lambda)$ to be very small by using the inequality $P(|Z| > x) \leq \exp(-x^2/2)$ for Z from $N(0,1)$ and for any $x > 0$. Contrary to this conventional setting, ours assumes x_t is stochastic, so that $\{x_{ti}u_t\}$ is no more independent Gaussian process. To evaluate the tail probability, we may use Azuma-Hoeffding's inequality together with the assumption that $\{x_{ti}u_t\}$ is a martingale difference sequence. In this case, we have to control the boundedness of $\{x_{ti}u_t\}$ at the same time, resulting in the additional factor $\log(pT)$ described above.

3.2 Oracle property

It is well known that the capacity of the Lasso for model selection is quite limited (e.g., Fan and Lv 2011). If we employ a SCAD-type penalty, however, a stronger and more desirable result on variable selection can be obtained. This result is called the oracle property, as studied first by Fan and Li (2001). The property admits $\hat{\boldsymbol{\beta}}_A$ to be asymptotically equivalent

to the maximum likelihood estimate obtained under the correct restriction $\boldsymbol{\beta}_B = \mathbf{0}$. To derive it under a time series setting, we need a different set of conditions; see Appendix A.1. Define $d(= d_T) \equiv \min_{j \in A} |\beta_{0,j}|/2$, $\boldsymbol{I}_{0AA} \equiv T\mathrm{E}[\boldsymbol{G}_{0AT}\boldsymbol{G}_{0AT}^\top]$, and $\boldsymbol{J}_{0AA} \equiv \mathrm{E}[\boldsymbol{H}_{AAT}]$.

Under assumptions in Appendix A.1, we will derive model selection consistency and appropriate rate of convergence. The role of Assumption 7 is essentially the same as that of Assumption 3. The first condition in Assumption 8 is a variant of the *beta-min condition* in Bülmann and van de Geer (2011, Ch. 7). This is necessary to distinguish the nonzero coefficient of relevant variables from zero though it seems stringent in the case of econometric modeling. The second condition $p'_\lambda(d) = 0$ is key to achieve the oracle property. This is strong enough to exclude the ℓ_1-penalty from Assumption 2. In fact, for the ℓ_1-penalty, $p'_\lambda(v) = \lambda(> 0)$ holds identically for all $v > 0$. On the other hand, for the SCAD and MCP, this holds for a sufficiently large T as long as $d/\lambda \to \infty$ is satisfied. Assumptions 9–11 seem quite natural and are frequently used in stationary time series analysis. Assumption 12 restricts the asymptotic behavior of the lower-left $(p - s) \times s$ submatrix of $\boldsymbol{H}_T$. This is essentially the same as condition (27) of Fan and Lv (2011).

Letting $\boldsymbol{b} \in \mathbb{R}^s$ be such that $\|\boldsymbol{b}\|_2^2 = 1$, we set $\xi_t \equiv \boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2} \boldsymbol{x}_{At} u_t$ and $\xi_{Tt} \equiv T^{-1/2}\xi_t$. These can easily be shown to be a martingale difference sequence and martingale difference array, respectively. Note that $\sum_{t=1}^T \xi_{Tt}$ can also be written as $T^{1/2}\boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2}\boldsymbol{G}_{0AT}$. Assumption 13 is required to obtain the asymptotic normality. From Davidson (1994, Ch. 24), this leads to a central limit theorem of a martingale difference sequence. If ξ_t is ergodic stationary, this is redundant (Billingsley, 1961).

Theorem 2 (oracle property) *Let Assumptions 1, 2, and 7–12 hold. Then, there exists a local minimizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_A^\top, \hat{\boldsymbol{\beta}}_B^\top)^\top$ *of* $Q_T(\boldsymbol{\beta})$ *such that*

(a) (Sparsity) $\hat{\boldsymbol{\beta}}_B = \mathbf{0}$ *with probability approaching one;*

(b) (Rate of convergence) $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_2 = O_p((s/T)^{1/2})$.

In addition, if Assumption 13 holds, then for any $\boldsymbol{b} \in \mathbb{R}^q$ *satisfying* $\|\boldsymbol{b}\|_2^2 = 1$, *we have*

(c) (Asymptotic normality) $T^{1/2}\boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2}\boldsymbol{H}_{AAT}^\top(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) \to_d N(0,1)$.

The oracle property means that the model selection is consistent in the sense of (a) and (b). Moreover, as is understood by result (c), the estimator has the same asymptotic efficiency as the (infeasible) MLE obtained with advance knowledge of the true submodel. Based on these results, we can estimate ultrahigh-dimensional models without irksome tests for zero restrictions on the parameters or an exhaustive search using information criteria.

4 Empirical Examples

According to the theoretical results given in the previous sections, the penalized regression can have two desirable properties: the oracle inequality and the oracle property. In this section, we provide two empirical examples that motivate how well the penalized regression works in macroeconometric analyses. The first forecasts the quarterly real U.S. GDP with a large number of monthly macroeconomic predictors, and the second screens portfolio from a large number of potential securities using NYSE stock price data.

4.1 Forecasting quarterly U.S. GDP with a large number of predictors

4.1.1 Penalized MIDAS regression model

In this section, we illustrate how to apply the penalized regression model to macroeconomic time series using the MIDAS forecasting regression. The MIDAS regression model was originally proposed by Ghysels et al. (2007) and is now one of standard tool for forecasting with mixed-frequency data, as well as the now-casting model based on the state-space representation (e.g., Giannone et al., 2008; Bańbura and Modugno, 2013). The original (or basic) MIDAS regression model has an advantage of describing a forecasting regression model in a simple and parsimonious way of a distributed lag structure with a few hyper-parameters. However, the original MIDAS regression model would not be suitable for a

situation where the number of predictors in the model is very large. For example, consider the original MIDAS regression model with K hyperparameters and N macroeconomic time series. Then, the total number of parameters in the original MIDAS regression model remains $NK + 1 = O(N)$. Thus, it invokes a serious efficiency loss if N is large or even it makes the model inestimable. On the other hand, the penalized regression enables us to estimate the MIDAS regression model without imposing the distributed lag structure on the regression coefficients. Moreover, Theorem 1 implies that the forecast value obtained by the penalized regression is reliable. In the following, we link the penalized regression model (1) to the MIDAS regression model without parameter restrictions, and consider to forecast quarterly U.S. GDP with the monthly macroeconomic data using the penalized regression.

Let $\{y_t, \boldsymbol{x}_{t/m}^{(m)}\}$ be the MIDAS process in line with Andreou et al. (2010), where the scalar y_t is the low-frequency variable observed at $t = 1, \ldots, T$, and the N-dimensional vector $\boldsymbol{x}_{t/m}^{(m)} = \left(1, x_{2,t/m}^{(m)}, \ldots, x_{N,t/m}^{(m)}\right)^\top$ is a set of higher-frequency variables observed m times between t and $t-1$. For example, $m = 3$ if we forecast a quarterly variable with monthly predictors. We consider the h-step-ahead mixed-frequency forecasting regression model with ℓ lags,

$$y_t = \boldsymbol{x}_{t-h}^\top \boldsymbol{\beta}_0 + u_t, \qquad t = 1, \ldots, T, \tag{4}$$

where $\boldsymbol{x}_{t-h} = \left(1, \boldsymbol{x}_{2,t-h,\ell}^{(m)}, \ldots, \boldsymbol{x}_{N,t-h,\ell}^{(m)}\right)^\top$ with $\boldsymbol{x}_{k,t-h,\ell}^{(m)} = \left(x_{k,t-h/m}^{(m)}, x_{k,t-h-1/m}^{(m)} \ldots, x_{k,t-h-\ell/m}^{(m)}\right)^\top$ for $k = 2, 3, \ldots, N$, $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,N\ell+N-\ell})^\top$ is the parameter vector and u_t is an error term. Here the case $h < 1$ ($h = 0, 1/m, 2/m, \ldots, (m-1)/m$) corresponds to nowcast; we forecast a low-frequency variable with the "latest" high-frequency variables released between $t-1$ and t. For instance, if we consider a quarterly/monthly ($m = 3$) case, $h = 0$ (1/3) means that we forecast a quarterly variable in 2015Q2 with monthly data in June (May) 2015 or later. Note that model (4) has the same structure as (1) with $p := (N-1)(\ell+1) + 1 = N\ell + N - \ell$ but it differs from the original MIDAS regression model by Ghysels et al. (2007); our model does not employ the distributed lag structure on $\boldsymbol{x}_{t-h}$ while they used $\boldsymbol{x}_{t-h}(\theta) = \left(1, x_{2,t-h}^{(m)}(\theta_2), \ldots, x_{N,t-h}^{(m)}(\theta_N)\right)^\top$ instead of $\boldsymbol{x}_{t-h}$ such that $x_{k,t}^{(m)}(\theta_k) = \sum_{j=1}^{\ell} w_{j,k}(\theta_k) L^{j/m} x_{k,t/m}^{(m)}$ for

$k = 2, \ldots, p$, where $w_{j,k}(\theta_k) \in (0, 1)$ and $\sum_{j=1}^{\ell} w_{j,k}(\theta_k) = 1$. As mentioned above, the original MIDAS model crucially depends on the restrictive distributed lag structure and cannot reduce the total number of the parameters to be estimated effectively if N is very large. Alternatively, the MIDAS regression that minimizes the penalized loss can estimate $\boldsymbol{\beta}_0$ and forecast y_t without the distributed lag structure.

In a macroeconomic forecasting point of view, it is natural to consider that there is a small set of key predictors that contain rich information to forecast y while there are lots of redundant predictors. To reduce accumulation of estimation errors, we should model y only by using the key predictors. Although the redundant predictors would have "non-zero" forecasting power, the penalized regression makes their coefficient estimates zero as an approximation. In other words, we can say that the sparsity assumption claims there exist "targeted predictors" for y (Bai and Ng, 2008).

Hereafter, we call the MIDAS regression model estimated by the penalized regression as "penalized MIDAS regression." We also note that as a method related to our penalized MIDAS regression, Marsilli (2014) proposes a MIDAS regression model with a penalized regression. However, he employed the original MIDAS parsimonious parameterization, which completely differs from our model in terms of parameterization as we stressed above.

4.1.2 Data

U.S. quarterly real GDP growth is taken from the FRED database. The sample period is from 1959Q4 to 2016Q2. We retrieve 117 U.S. monthly macroeconomic time series ($N = 117$) from the FRED–MD database and the series are appropriately detrended according to a guideline given in McCracken and Ng (2015). Note that the FRED–MD database originally contains a total of 128 series, but we remove 11 series due to the following reasons: the CBOE S&P 100 Volatility Index (VXOCLSx), Consumer sentiment index (UMCSENTx), Trade weighted U.S. dollar index of major currencies (TWEXMMTH), New orders for nondefense capital goods (ANDENOx), New orders for consumer goods

(ACOGNO), and New private housing permits (PERMIT, PERMITNE, PERMITMW, PERMITS, PERMITW) have no observations from 1959. Furthermore, our preliminary inspection found that Reserves of depository institutions nonborrowed (NONBORRES) contained extreme changes in February 2008, which would contaminate our analysis. The sample period of the detrended monthly series is from March 1959 (1959:3) to June 2016 (2016:6).

4.1.3 Forecasting Strategy

We evaluate the out-of-sample forecasting performance by mean squared forecast errors (MSFE) in the evaluation period from 2000Q1 to 2016Q2. The parameter estimates are obtained from each estimation period; the initial period is 1959Q4–1999Q4 and the next one extends the end point to 2000Q1 with the starting point 1959Q4 being fixed. For example, the initial forecast error in 2000Q1 is calculated using the estimates from the initial estimation period 1959Q4–1999Q4, and the second forecast error in 2000Q2 uses the estimates from the second estimation period 1959Q4–2000Q1. We suppose that the forecast regression consists of eight lags ($\ell = 8$), so that the total number of parameters for the forecasting regression to be estimated is $N\ell + N - \ell = 117 \times 8 + 117 - 8 = 1045$, including a constant term. The penalized MIDAS regression is expected to be robust to a choice of ℓ, as long as we choose ℓ to be moderately large, because the penalized regression conducts model selection as well as parameter estimation. To investigate the forecasting performance of the penalized MIDAS regression model with a variety of horizons, we examine cases where $h = 0, 1/3, 2/3, 1, 4/3, 5/3, 2$ in the same manner as Clements and Galvão (2008) and Marcellino and Schumacher (2010). The cases $h = 0, 1/3$, and $2/3$ correspond to nowcasting in the sense that we forecast contemporaneous or very short-forecast-horizon quarterly GDP growth using monthly series before the official announcement of the GDP, while the case $h = 2$ is a forecast with a relatively long horizon. The sample size of the estimation period T gradually increases and varies depending on h; for example, T ranges from 161 to 227 if $h = 0$, and from 159 to 225 if $h = 2$.

Finally, we need to determine the values of the tuning parameters, a and λ, in advance of the penalized MIDAS regression regression. Following the guidelines by Breheny and Huang (2011, pp. 19 and 21) with our preliminary inspection of the overall samples, we set $a = 12$ for the SCAD and MCP, although the performance could be improved by a more careful choice. The value of λ is selected by 10-fold cross-validation. The validity was confirmed by Uematsu and Tanaka (2015). All estimations for the penalized regression are conducted using R 3.2.1 with the `ncvreg` package of Breheny and Huang (2011).

4.1.4 Forecast performance

To measure the performance appropriately, we consider two types of datasets. The first is a complete dataset, that is, there are no missing values in the dataset. The second is a real-time dataset, which has jagged/ragged edge pattern due to the publication lag of the series.

4.1.5 Forecast performance in complete data

We use data from 1959Q4–2016Q1 for the GDP and 1959:3 to 2016:3 to retrieve a complete dataset. We consider the following three evaluation periods: (*i*) Overall (2000Q1–2016Q1), (*ii*) 1st subsample (2000Q1–2007Q4), and (*iii*) 2nd subsample (2008Q1–2016Q1). This is because the unprecedented turmoil of the U.S. economy stemming from the subprime mortgage crisis and the ensuing collapse of Lehman Brothers in 2008 would introduce parameter instability that would distort the forecast evaluation. As a result, we consider the forecast performance of the penalized regression in complete data from a total of 65 (overall), 32 (1st subsample) and 33 (2nd subsample) squared forecast errors, respectively.

Tables 1–3 report the mean squared forecast errors (MSFE) of the penalized MIDAS regression with the SCAD, MCP, and Lasso, and their two competitors in the overall sample, 1st subsample, and 2nd subsample, respectively. In the tables, the median squared forecast errors are also shown in parentheses to remove contamination by outliers. The all values are relative values compared to a naive AR(4) forecast. The two competitors are the factor

MIDAS (denoted "Factor" in the tables) proposed by Marcellino and Schumacher (2010) and the two-step penalized regression (post-OLS) procedure (denoted as "post-MCP," "post-SCAD," "post-Lasso" in the tables) proposed by Belloni and Chernozhukov (2013). The factor MIDAS is expected to be one of the strong competitors since the factor-based forecast is found to perform well in forecasting real variables (e.g., Stock and Watson, 2002, 2012; De Mol et al., 2008.). The factor MIDAS considered here is based on the basic MIDAS structure with the exponential Almon lag structure of two hyperparameters. The number of factors is assumed to be seven ($r = 7$) based on the information criterion IC_{p2} by Bai and Ng (2002). Although we can consider the *unrestricted* Factor MIDAS as in Marcellino and Schumacher (2010), which is free from the distributed lag structure, we do not employ it because of its intractability caused by high dimensionality. The two-step procedure using the Lasso is known as the *OLS post-Lasso*. Belloni and Chernozhukov (2013) showed it could perform at least as well as the Lasso and could be better in some cases. We also consider the two-step procedure using the MCP and SCAD penalties.

First, we consider the nowcasting ($0 \leq h < 1$) cases. Table 1 shows that all methods are much better than the naive AR(4) forecast, but that the penalized MIDAS regression outperforms the factor MIDAS and the two-step procedures in the overall sample with a few exceptions, in terms of both the mean and median squared forecast errors. The two-step procedures work well in terms of MSFE, but do not seem to work well in the median measure since they are frequently beaten by the naive AR(4) forecast. We also find that the MSFE of the factor when $h = 1/3$ is much worse than other methods, owing to outliers of forecast values around the subprime mortgage crisis. Tables 2 and 3 show the forecasting performance for the first and second subsamples, respectively. In first subsample, the penalized MIDAS regression does not necessarily work well; it performs well when $h = 0$, but worse than the factor MIDAS when $h = 1/3$ and $2/3$. However, we also find that the penalized MIDAS regression performs well and completely dominates the factor MIDAS and the two-step procedures in the second subsample in terms of both mean and median measures. Thus, it can

be said that the penalized MIDAS regression is more robust than the other methods in terms of structural instability. Furthermore, we find that the MSEs of the two-step procedure are worse than those of the penalized MIDAS regression, overall. Thus, the two-step procedure does not provide effective efficiency gains in our situation. A probable reason is that the total number of regressors in the second-step OLS regression does not become effectively small when we assume a long-length lag structure in the model even if variable "screening" is conducted in the first step. This would make the efficiency losses arising from estimating many parameters more serious than estimating penalized MIDAS regression directly. Next, we turn to the forecast performance when $h \geq 1$. The tables show that all the methods have similar forecast performances; they perform well when $h = 1$, however, when $h > 1$, they are all beaten by the AR(4) forecast. The results are not surprising because Clements and Galvão (2008) and Marcellino and Schumacher (2010) also find the same results. Hence, our results show that the penalized MIDAS has a good forecast performance in a very short horizon, especially in the presence of instability, although it is not necessarily a primary tool for a forecast with a relatively long horizon. However, we can conclude that penalized MIDAS regression is an effective tool for forecasting with mixed-frequency data because our main interest in forecasting with mixed-frequency data is nowcasting where low-frequency data are not available.

4.1.6 Forecast performance in real-time data

Section 4.1.5 reveals that the penalized regression behaves well in nowcasting with a complete data. However, when we actually conduct real-time forecasting of quarterly GDP with monthly data, a complete dataset is not available because of possible publication lags of the series. Thus, we must face an incomplete dataset so called "jagged (ragged)-edge" dataset, that contains missing values in some latest months. Then we investigate how well the forecast with penalized regression works with the real-time data. It should be mentioned that in our experiment, strictly speaking, we consider "pseudo" real-time forecasting; we suppose

Table 1: Mean/Median Forecast Errors of the forecasts in complete data [Overall Sample]

	$h=0$	$h=1/3$	$h=2/3$	$h=1$	$h=4/3$	$h=5/3$	$h=2$
MCP	0.58	0.50	0.53	0.80	1.17	1.35	1.34
(median)	(0.80)	(0.79)	(0.80)	(0.57)	(1.32)	(1.16)	(1.18)
SCAD	0.59	0.53	0.61	0.79	1.15	1.34	1.34
(median)	(0.76)	(0.86)	(0.75)	(0.60)	(1.28)	(1.20)	(1.19)
Lasso	0.56	0.56	0.60	0.79	1.15	1.33	1.34
(median)	(0.80)	(0.89)	(0.70)	(0.61)	(1.27)	(1.17)	(1.30)
Factor	0.83	2.12	0.89	0.75	1.89	1.25	1.25
(median)	(1.11)	(0.89)	(0.81)	(1.00)	(1.12)	(1.62)	(1.73)
post-MCP	0.79	0.62	0.61	0.82	1.16	1.43	1.50
(median)	(1.48)	(1.19)	(1.01)	(0.79)	(1.09)	(1.63)	(2.06)
post-SCAD	0.79	0.60	0.56	0.81	1.21	1.35	1.46
(median)	(1.20)	(1.27)	(0.81)	(0.88)	(1.17)	(1.41)	(1.59)
post-Lasso	0.82	0.61	0.93	0.81	1.17	1.36	1.46
(median)	(1.35)	(1.14)	(0.92)	(0.88)	(1.02)	(1.48)	(1.55)

Note) All values are relative values to AR(4) forecast. Values in parentheses are median forecast errors.

each monthly data for all evaluation periods have the same jagged (ragged)-edge pattern as of the 2016-08 version of the FRED-MD. For example, Real manufacturing and trade industry sales (CMRMTSPLx) and the Help-wanted index (HWI) have one and four month missing values owing to publication lags in the 2016-08 version, respectively. Then we suppose the data for all estimation periods have the same jagged-edge patterns even if our dataset contains complete data for those periods. Moreover, we assume no data revisions occur in our dataset.

Tables 4–6 show the relative MSFEs of the penalized regression and the state-space ML estimator proposed by Bańbura and Modugno (2014) in the real-time overall sample (2000Q1–2016Q2), 1st subsample (2000Q1–2007Q4), and 2nd subsample (2008Q1–2016Q2), respectively. The tables omit the results for $h > 1$ and concentrate on the nowcast

Table 2: Mean/Median Forecast Errors of the forecasts in complete data [1st Subsample]

	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$	$h = 4/3$	$h = 5/3$	$h = 2$
MCP	0.76	0.74	0.70	0.93	1.03	1.29	0.76
(median)	(0.84)	(0.94)	(1.15)	(0.90)	(1.37)	(1.77)	(0.84)
SCAD	0.77	0.78	0.71	0.92	1.03	1.27	0.77
(median)	(0.81)	(0.97)	(1.07)	(0.83)	(1.28)	(1.53)	(0.81)
Lasso	0.74	0.77	0.76	0.92	1.03	1.27	0.74
(median)	(0.59)	(0.88)	(1.15)	(0.84)	(1.28)	(1.51)	(0.59)
Factor	0.86	0.69	0.60	0.86	1.06	1.76	0.86
(median)	(0.98)	(0.86)	(0.86)	(1.27)	(1.46)	(3.34)	(0.98)
post-MCP	0.95	1.09	0.92	1.09	1.23	1.60	0.95
(median)	(1.52)	(1.31)	(1.52)	(1.28)	(1.51)	(2.41)	(1.52)
post-SCAD	1.21	1.00	0.76	1.04	1.07	1.55	1.21
(median)	(0.75)	(1.22)	(1.06)	(1.59)	(1.69)	(1.52)	(0.75)
post-Lasso	1.33	1.03	1.51	1.07	1.06	1.60	1.33
(median)	(1.19)	(1.40)	(1.54)	(1.60)	(1.35)	(1.52)	(1.19)

Note) All values are relative values to AR(4) forecast. Values in parentheses are median forecast errors.

situation ($0 \leq h \leq 1$) because the real-time forecasting is meaningful only in a very short horizon. The state-space ML estimation enables us to handle real-time mixed frequency data by embedding missing patterns of data in the model; see Bańbura and Modugno (2014) for details. On the other hand, the penalized regression requires an interpolated dataset to obtain the forecast values. Thus, we employ an interpolation method based on the EM algorithm proposed by Stock and Watson (2002).

From the tables, we first find the effects of the jagged-edge and interpolation on the forecast accuracy of the penalized regression are negligible since they do not essentially affect the mean/median squared forecast errors values compared with the results in Tables 1–3. Second, we see that the penalized regression performs well in the overall and 2nd subsample; it beats the state-space ML when $h = 2/3$ and 1 in both the mean/median measures, and

Table 3: Mean/Median Forecast Errors of the forecasts in complete data [2nd Subsample]

	$h=0$	$h=1/3$	$h=2/3$	$h=1$	$h=4/3$	$h=5/3$	$h=2$
MCP	0.49	0.38	0.45	0.73	1.24	1.38	1.37
(median)	(0.86)	(0.79)	(0.79)	(0.54)	(1.32)	(0.96)	(1.16)
SCAD	0.50	0.41	0.57	0.73	1.21	1.37	1.38
(median)	(0.76)	(0.89)	(0.66)	(0.56)	(1.35)	(1.20)	(1.30)
Lasso	0.48	0.47	0.52	0.73	1.21	1.37	1.38
(median)	(1.09)	(1.06)	(0.68)	(0.59)	(1.35)	(1.15)	(1.30)
Factor	0.82	2.81	1.03	0.69	2.30	0.99	1.19
(median)	(1.54)	(1.39)	(0.94)	(1.07)	(1.09)	(1.53)	(1.43)
post-MCP	0.71	0.40	0.46	0.69	1.13	1.35	1.38
(median)	(1.56)	(1.28)	(0.84)	(0.72)	(0.89)	(1.56)	(2.24)
post-SCAD	0.59	0.40	0.47	0.70	1.28	1.25	1.36
(median)	(1.78)	(1.48)	(0.82)	(0.72)	(1.14)	(1.60)	(1.54)
post-Lasso	0.57	0.41	0.65	0.68	1.23	1.24	1.32
(median)	(1.65)	(1.09)	(0.77)	(0.72)	(1.14)	(1.60)	(1.46)

Note) All values are relative values to AR(4) forecast. Values in parentheses are median forecast errors.

Table 4: Mean/Median Forecast Errors of the forecasts in jagged-edge data [Overall sample]

	$h=0$	$h=1/3$	$h=2/3$	$h=1$
MCP	0.58	0.50	0.54	0.80
(median)	(0.82)	(0.81)	(0.88)	(0.62)
SCAD	0.60	0.53	0.62	0.80
(median)	(0.79)	(0.88)	(0.81)	(0.67)
Lasso	0.57	0.57	0.61	0.80
(median)	(0.92)	(0.93)	(0.78)	(0.67)
State-Space ML	0.47	0.49	0.66	0.84
(median)	(0.70)	(0.71)	(1.00)	(1.02)

Note) All values are relative values to AR(4) forecast. Values in parentheses are median forecast errors.

performs as well as the state-space ML when $h = 1/3$ while it does not relatively work well in the 1st subsample as in the complete data case. The state-space ML is expected to have higher forecasting performance than the penalized regression because the state-space ML is based on a system equation with richer information while the penalized regression relies on a single equation. However, this would not be true when a model misspecification is present, as Bai et al. (2013) claimed. Then, our results that reveal the penalized regression can be compete with the state-space ML in terms of forecasting accuracy imply that the system equation contains a certain level of the misspecification. Moreover, it should be mentioned that the penalized regression is much simpler and rapid than the state-space ML in obtaining the forecast values. Since the dimension of the state-space model can be very large when we forecast with mixed frequency (117 dimensional state-space models with 40 latent factors in our case), the estimation is much computationally demanding and time consuming (roughly eight times longer than the penalized regression). Furthermore, the estimated values can be unstable if we consider to apply the state-space ML to a dataset with larger N and/or r.

Although we do not examine them here, the Ridge regression and the Bayesian VAR (BVAR) would be potential alternatives to the state-space ML (e.g., De Mol et al., 2008; and Schorfheide and Song, 2015). However, they are also computationally demanding (the BVAR requires more than 100,000 parameter estimation in our case) and their theoretical properties have not been investigated yet under "ultra"high- dimensionality (i.e. p diverges at a sub-exponential rate).

4.2 Screening Effective Portfolio from a Large Number of Potential Securities

Recent studies on portfolio selection have focused on the penalized regression because it plays a crucial role in constructing a portfolio when there are a large number of potential stocks. Brodie et al. (2009) find out the penalized regression is useful in selecting optimal portfolio in terms of the out-of-sample performance measured by the Sharpe ratio; Fan et

Table 5: Mean/Median Forecast Errors of the forecasts in jagged-edge data [1st Subsample]

	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$
MCP	0.76	0.74	0.70	0.93
(median)	(0.84)	(0.94)	(1.15)	(0.90)
SCAD	0.77	0.78	0.71	0.92
(median)	(0.81)	(0.97)	(1.07)	(0.83)
Lasso	0.74	0.77	0.76	0.92
(median)	(0.59)	(0.88)	(1.15)	(0.84)
State-Space ML	0.62	0.67	0.71	0.94
(median)	(0.72)	(0.72)	(0.65)	(1.17)

Note) All values are relative values to AR(4) forecast. Values in parentheses are median forecast errors.

al. (2012) introduced gross-exposure constraints to admit short sales in the estimation of an optimal portfolio; Carrasco and Noumon (2012) focused on estimating a precision matrix of returns. They found the penalized regression is quite useful to stabilize the estimation of the covariance matrix and provided better finite sample performances than traditional methods.

To the best of our knowledge, the existing literature concerning applications of the penalized regression to portfolio selection focused on yieldability. However, it seems interesting to examine the consistent estimation of weights of the portfolio; that is, screening how fund managers construct their portfolio from a large number of securities is valuable. Unlike the other high-dimensional estimation methods, such as the factor and the Ridge, the SCAD-type penalized regression enables us to screen their portfolio from a large dataset of stock prices. In this section, we examine how well the penalized regression usefully works in this direction using a large NYSE stock price dataset.

4.2.1 Construction of Portfolio

Suppose a fund manager faces p potential stocks, where x_{it} is the rate of return of the ith ($i = 1, 2, \ldots, p$) stock at time t. Let $\boldsymbol{x}_t = [x_{1t}, x_{2t}, \ldots, x_{pt}]^\top$ be the p-dimensional rate of the

Table 6: Mean/Median Forecast Errors of the forecasts in jagged-edge data [2nd Subsample]

	$h = 0$	$h = 1/3$	$h = 2/3$	$h = 1$
MCP	0.50	0.39	0.46	0.74
(median)	(1.06)	(0.82)	(0.81)	(0.55)
SCAD	0.51	0.42	0.58	0.74
(median)	(0.83)	(0.90)	(0.71)	(0.65)
Lasso	0.49	0.48	0.53	0.74
(median)	(1.13)	(1.10)	(0.69)	(0.65)
State-Space ML	0.41	0.40	0.63	0.79
(median)	(0.84)	(0.76)	(1.05)	(1.12)

Note) All values are relative values to AR(4) forecast. Values in parentheses are median forecast errors.

return vector at t and $\boldsymbol{\omega}_0$ be the p-dimensional weight vector of the portfolio that satisfies $\|\boldsymbol{\omega}_0\|_0 = s \ (\ll p)$, $\boldsymbol{\iota}'\boldsymbol{\omega}_0 = 1$ and $\|\boldsymbol{\omega}_0\|_1 = \zeta_w$, where $\zeta_w \in [1, \infty)$ and $\boldsymbol{\iota}$ is a p-dimensional vector with all elements being one. That is, the portfolio is constructed by s stocks from p potential stocks. We assume the fund manager constructs her portfolio as

$$y_t = \boldsymbol{x}_t^\top \boldsymbol{\omega}_0 + u_t, \qquad t = 1, \ldots, T, \tag{5}$$

where u_t is a "miscellaneous" component that includes all assets in the portfolio other than stocks, such as T-bills and corporate bonds. Further we assume that $\boldsymbol{x}_t$ and u_t are independent of each other and $u_t \sim i.i.d.N(0, \sigma_u^2)$, where $\sigma_u^2 = T^{-1}\boldsymbol{\omega}_{0A}^\top \boldsymbol{X}_A^\top \boldsymbol{X}_A \boldsymbol{\omega}_{0A}/\text{SNR}$, $\boldsymbol{\omega}_{0A}$ is a nonzero s-dimensional subvector of $\boldsymbol{\omega}_0$, $\boldsymbol{X}_A$ is $T \times s$ submatrix of $\boldsymbol{X}$ that corresponds to $\boldsymbol{\omega}_{0A}$, and $\text{SNR} = V(\boldsymbol{x}_t^\top \boldsymbol{\omega}_0)/V(u_t)$. Although we might consider the case in which $\boldsymbol{x}_t$ and u_t are dependent by extending the results of Fan and Liao (2014), this is beyond the scope of our research, and we regard $\boldsymbol{x}_t$ and u_t as independent here.

The portfolio allows short sales if $\zeta_w > 1$ with ζ_w determining a constraint on the short sales as shown in Fan et al. (2012). Let $w_0^+ = (\zeta_w + 1)/2$ and $w_0^- = (\zeta_w - 1)/2$. Then w_0^+ and w_0^- correspond to the total proportions of long and short sales, respectively, since $w_0^+ + w_0^- = \zeta_w = \|\boldsymbol{\omega}_0\|_1$ and $w_0^+ - w_0^- = 1$, and w_0^- becomes larger as ζ_w grows while short

sales are not allowed if $\zeta_w = 1$ ($w_0^- = 0$). We assume the fund manager holds equal amounts of long and short sales of $s/2$ and that she employs equal weights among long and short sales; that is, we assume $\omega_{0i} = w_0^+/(s/2)$ for $i \in \boldsymbol{\omega}_{0A+}$, $-w_0^-/(s/2)$ for $i \in \boldsymbol{\omega}_{0A-}$, and 0 for $i \in \boldsymbol{\omega}_{0B}$, where ω_{0i} is ith element of $\boldsymbol{\omega}_0$, and $\boldsymbol{\omega}_{0A+}$, $\boldsymbol{\omega}_{0A-}$, and $\boldsymbol{\omega}_{0B}$ are sets of stocks of long, short, and no sales, respectively.

4.2.2 Data and Evaluation Strategy

We retrieve weekly stock price data of the NYSE from *Yahoo! Finance*. Our dataset contains 1853 adjusted stock prices ($p = 1853$) with starting from the 1st week of January in 2009 to the 4th week of April in 2016. In this application, we apply the log-difference to the stock price data and standardize them so that the data are converted to rates of returns with zero means and unit variances. We investigate the cases of $s = 34$ and 40 with $a = 14$, SNR = 10, and $\zeta_w = 10$. Non zero s stocks are drawn randomly from p candidates with equal probabilities. Furthermore, we assume the fund manager does not rebalance the portfolio. Hence it remain unchanged in all sample period. Brodie et al. (2009) argue a possibility of estimating a weight vector for a portfolio in the presence of rebalancing with a penalized regression, but we do not consider the case here.

The purpose of this application is to screen the kinds of stocks in which the fund manager invests from a large number of potential stocks using the penalized regression. We examine how well the penalized estimator $\hat{\boldsymbol{\omega}}$ can distinguish the nonzeros from zero elements of $\boldsymbol{\omega}_0$ in finite samples. Then we evaluate the finite sample properties of $\hat{\boldsymbol{\omega}}$ to focus on SC-A $= P\left(\text{sgn}(\hat{\boldsymbol{\omega}}_A) = \text{sgn}(\boldsymbol{\omega}_{0A})\right)$ and SC-$B = P\left(\text{sgn}(\hat{\boldsymbol{\omega}}_B) = \text{sgn}(\boldsymbol{\omega}_{0B})\right)$; the SC-$A$ is the success rate of detecting non-zero elements of $\boldsymbol{\omega}_0$ with the correct sign and the SC-B is that of detecting zero elements. We expect that the SCAD-type penalized regression estimator can have high SC-A and SC-B values as T becomes large thanks to the oracle property. The SC-A and SC-B are sequentially computed for 172 evaluation periods in this application, where the endpoint gradually grows by one while the start point is fixed; the initial evaluation period starts from

the 2nd week of January 2009 and ends in 1st week of December 2010 ($T = 209$). The 2nd evaluation period runs from the 2nd week of January 2009 to the 2nd week of December 2010 ($T = 210$), and so on. The terminal evaluation period is from the 2nd week of January 2009 to the 4th week of April 2016 ($T = 381$).

4.2.3 Empirical Results

Figures 2–3 and Figures 4–5 show the SC-A and SC-B of the MCP, SCAD, and Lasso for 172 evaluation periods with $s = 34$ and 40, respectively. To begin with, we consider the SC-A. At a glance, both Figures 2 and 3 reveal two characteristics of $\hat{\omega}$. First, the SC-A increases toward 1 as T grows for all penalties. Although the SC-A of $s = 40$ seems uniformly lower than that of $s = 34$ for all T, this is due to the fact that more nonzero elements requires a greater search cost. Second, the SC-A of the Lasso tends to be higher than that of the MCP and SCAD when T is relatively small, while it seems reversed when T grows large. This is consistent with the theory because the Lasso tends to have many "false positive" estimates. That is, it overestimates the total number of nonzero elements since it rarely satisfies the assumptions for model selection consistency, while the MCP and SCAD satisfy these assumptions in many cases, as argued in Appendix A.2. Then, the SC-A of the Lasso is not expected to be higher than that of the MCP and SCAD when T is large.

Next, we focus on the SC-B. Figures 4 and 5 show that SC-B of the MCP and SCAD are successfully nearly equal to 1 and dominate that of the Lasso for all T. The results are consistent with the theory because the MCP and SCAD have the oracle property, which means they can detect true zero parameters more precisely than the Lasso can, except for extraordinary cases.

In summary, our empirical results reveal that the model selection consistency of the SCAD-type penalty works well in a large stock price dataset. This implies that the penalized regression enables us to effectively detect the behavior of fund managers from large financial datasets.

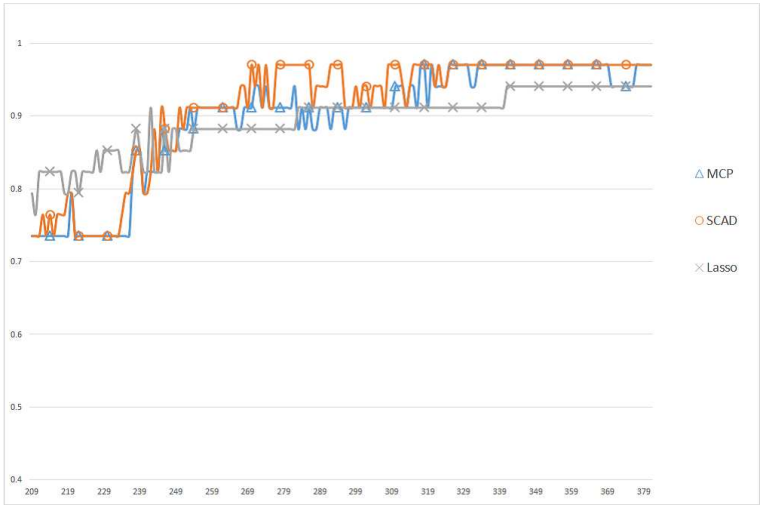

Figure 2: SC-A when $s = 34$ (from $T = 209$ to $T = 381$)

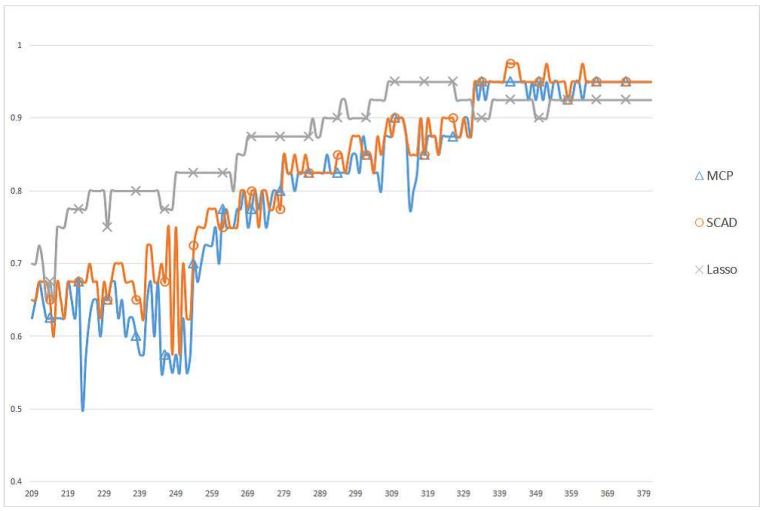

Figure 3: SC-A when $s = 40$ (from $T = 209$ to $T = 381$)

5 Conclusion

We have studied macroeconomic forecasting and variable selection using a folded-concave penalized regression with a very large number of predictors. The contributions include both theoretical and empirical results. The first half of the paper developed the theory for a folded-concave penalized regression in ultrahigh dimensions when the model exhibits time

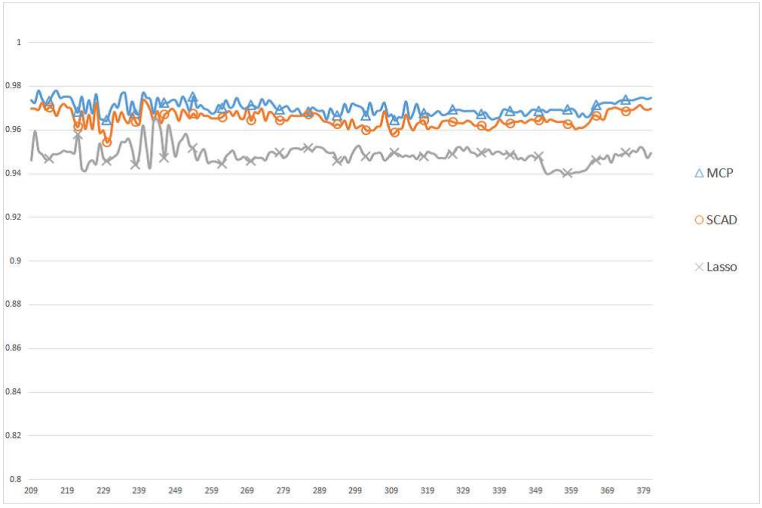

Figure 4: SC-B when $s = 34$ (from $T = 209$ to $T = 381$)

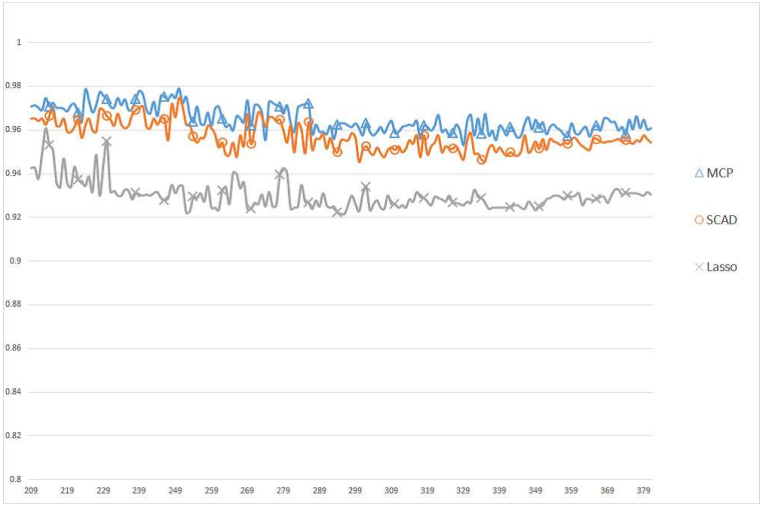

Figure 5: SC-B when $s = 40$ (from $T = 209$ to $T = 381$)

series dependences. Specifically, we have proved the oracle inequality and the oracle property under appropriate conditions for macroeconomic time series. The latter half of the paper provided two empirical applications that motivated us to use the penalized regression for a large macroeconomic dataset. The first was the forecasting of quarterly U.S. real GDP with a large amount of monthly macroeconomic data taken from the FRED-MD through the MIDAS regression framework; the forecasting model consisted of more than 1000 monthly

predictors including lags while the sample size was much smaller than the total number of predictors. The forecasting performance of the penalized regression is promising one compared to that of the factor MIDAS proposed by Marcellino and Schumacher (2010) and the state-space (nowcasting) model of Bańbura and Modugno (2013). The second application screened a portfolio that contained about 40 stocks from more than 1800 stocks using NYSE stock price data. The oracle property ensured the variable selection consistency, that is, the penalized regression with the SCAD-type penalty could detect the portfolio from the data theoretically. In fact, we observed that the variable selection consistency worked properly when screening the portfolio. Our theoretical and empirical contributions are expected to introduce econometricians to the world of ultrahigh dimensional macroeconomic data.

Acknowledgement

The authors are grateful for the invaluable comments of the associate editor, the anonymous referees, Jinchi Lv, Ryo Okui, Yohei Yamamoto, and the participants of the econometric workshop in Kyoto University. All remaining errors are ours. The authors also thank Michelle Modugno for providing MATLAB codes used in Bańbura and Modugno (2013), which were helpful when coding our R codes. Uematsu acknowledges the financial support from a Grant-in-Aid for JSPS Fellows No.26-1905. Tanaka acknowledges the financial supports from a JSPS Grant-in-Aid for Young Scientists (B) No.16K17100 and the Joint Usage and Research Center, Institute of Economic Research, Hitotsubashi University.

References

[1] Ahn, S. C., and A. R. Horenstein (2013), "Eigenvalue Test for the Number of Factors," *Econometrica*, **81**, 1203–1227.

[2] Andreou, E., E. Ghysels and A. Kourtellos (2010), "Regression models with mixed sampling frequencies," *Journal of Econometrics*, **158**, 246–261.

[3] Bai, J., E. Ghysels, and J. H. Wright (2013), "State Space Models and MIDAS Regressions," *Econometric Reviews* , **32**, 779–813.

[4] Bai, J. and S. Ng (2002), "Determining the number of factors in approximate factor models," *Econometrica*, **70**, 191–221.

[5] Bai, J. and S. Ng (2008), "Forecasting economic time series using targeted predictors," *Journal of Econometrics*, **146**, 304–317.

[6] Bańbura, M., D. Giannone, M. Modugno and L. Reichlin (2013), "Now-casting and the real-time data flow," *Working Paper Series*, No.1564, European Central Bank.

[7] Bańbura and M. Modugno (2013), "Maximum Likelihood Estimation of Factor Models on Datasets with Arbitrary Pattern of Missing Data," *Journal of Applied Econometrics*, **29**, 133–160.

[8] Basu, S. and Michailidis (2015), 'Regularized Estimation in Sparse High-Dimensional Time Series" *Annals of Statistics*, **43**, 1535–1567.

[9] Belloni, A. and V. Chernozhukov (2013), "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, **19**, 521–547.

[10] Bickel P. J., Y. Ritov and A. B. Tsybakov (2010), "Hierarchical selection of variables in sparse high-dimensional regression," *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown.* Institute of Mathematical Statistics, 56–69.

[11] Billingsley, P. (1961), "The Lindeberg-Levy theorem for martingales," The Proceeding of American Mathematical Society, 788-792.

[12] Breheny, P. and J. Huang (2011), "Coordinate descent algorithm for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics*, **5**, 232–253.

[13] Brodie, J., I. Daubechies, C. De Mol, D. Giannone, and I. Loris (2009), "Sparse and Stable Markowitz Portfolios," *Proceedings of the National Academy of Sciences*, **106**, 12267–12272.

[14] Bühlmann, P. and S. van de Geer (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer.

[15] Carrasco, M. and N. Noumon (2012), "Optimal Portfolio Selection using Regularization," mimeo.

[16] Clements, M. and A. B. Galvão (2008), 'Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States," *Journal of Business and Economic Statistics*, **26**, 546–554.

[17] Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press.

[18] De Mol, C., D. Giannone and L. Reichlin (2008), "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics*, **146**, 318–328.

[19] Fan, J. and R. Li (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, **96**, 1348–1360.

[20] Fan, J. and J. Lv (2011), "Nonconcave penalized likelihood with NP-dimensionality," *IEEE Transactions on Information Theory*, **57**, 5467–5484.

[21] Fan, J., J. Zhang and K. Yu (2012), "Vast Portfolio Selection with Gross-Exposure Constraints," *Journal of the American Statistical Association*, **107**, 592–606.

[22] Fan, Y. and J. Lv (2013), "Asymptotic equivalence of regularization methods in thresholded parameter space," *Journal of the American Statistical Association*, **108**, 1044–1061.

[23] Giannone, D., L. Reichlin and D. Small (2008), "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, **55**, 665–676.

[24] Ghysels, E., A. Sinko and R. Valkanov (2007), "MIDAS regressions: Further results and new directions," *Econometric Reviews*, **26**, 53–90.

[25] Kock, A. B. and L. Callot (2015), "oracle Inequalities for high dimensional vector autoregressions," *Journal of Econometrics*, **186**, 325–344.

[26] Loh P.-L. and M. J. Wainwright (2014), "Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima," *Journal of Machine Learning Research*, **16**, 559–616.

[27] Lv, J. and Y. Fan (2009), "A unified approach to model selection and sparse recovery using regularized least squares," *Annals of Statistics*, **37**, 3498–3528.

[28] Marcellino, M. H. and C. Schumacher (2010), "Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP," *Oxford Bulletin of Economics and Statistics*, **72**, 518–550.

[29] Marsilli, C. (2014),"Variable selection in predictive MIDAS models," *Banque de France Working Paper*, **520**.

[30] McCracken, M. W. and S. Ng (2015),"FRED-MD: A monthly database for macroeconomic research," *Federal Reserve Bank of ST. Louis Working Paper Series*, **2015-012A**.

[31] Nardi, Y. and A. Rinaldo (2011), "Autoregressive process modeling via the lasso procedure," *Journal of Multivariate Analysis*, **102**, 528–549.

[32] Negahban, S. N., P. Ravikumar, M. J. Wainwright and B. Yu (2012),"A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers,"*Statistical Science*, **27**, 538–557.

[33] Nicholson, W. B., D. S. Matteson, and J. Bien (2015), "VARX-L: Structured Regularization for Large Vector Autoregressions with Exogenous Variables," *arXiv:1508.07497*.

[34] Schorfheide, F. and D. Song (2015), "Real-Time Forecasting With a Mixed-Frequency VAR," *Journal of Business & Economic Statistics*, **33**, 366–380.

[35] Song, S. and P. J. Bickel (2011), "Large vector auto regressions," *arXiv:1106.3915v1*.

[36] Stock, J. H. and M. W. Watson (2002), "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, **97**, 1167–1179.

[37] Stock, J. H. and M. W. Watson (2012), "Generalized shrinkage methods for forecasting using many predictors," *Journal of Business & Economic Statistics*, **30**, 481–493.

[38] Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of Royal Statistical Society Series B*, **58**, 267–288.

[39] Uematsu, Y. and S. Tanaka (2015), "Regularization Parameter Selection via Cross-Validation in the Presence of Dependent Regressors: A Simulation Study," *Economics Bulletin*, **36**, 313–319.

[40] Wainwright, M. J. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso),"

[41] Wang, H., G. Li and C.–L. Tsai (2007), "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *Journal of Royal Statistical Society, Series B*, **69**, 63–78.

[42] Zhang, C.–H. (2010), "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, **38**, 894–9421.

Appendix

A.1 Assumptions for the oracle property

Assumption 7 There is a sequence $\lambda = o(1)$ such that $\|\boldsymbol{G}_{0BT}\|_\infty \leq \lambda/2$ holds with high probability.

Assumption 8 $(s/T)^{1/2} \ll \lambda \ll d = o(1)$ and $p'_\lambda(d) = 0$ for a sufficiently large T.

Assumption 9 For all i, $\max_t \mathrm{E}\left[(x_{ti}u_t)^2\right] < \infty$.

Assumption 10 There exists a constant c_H such that the Hessian submatrix satisfies with high probability, $\min_{\boldsymbol{v}\in\mathbb{R}^s} \boldsymbol{v}^\top \boldsymbol{H}_{AAT}\boldsymbol{v} \geq c_H\|\boldsymbol{v}\|_2^2$.

Assumption 11 $c_I \leq \Lambda_{\min}(\boldsymbol{I}_{0AA}) \leq \Lambda_{\max}(\boldsymbol{I}_{0AA}) \leq 1/c_I$ for a (small) constant $c_I > 0$.

Assumption 12 $\|\boldsymbol{H}_{BAT}\|_{2,\infty} \equiv \max_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{H}_{BAT}\boldsymbol{v}\|_\infty = O_p(1)$.

Assumption 13 $\mathrm{E}|\xi_t|^{2+\delta} \leq c_\xi$ for some constant $c_\xi > 0$.

A.2 Model selection inconsistency of Lasso

As far as forecasting is concerned, Theorem 1 shows that the resulting performance does not depend on the choice of penalties. However, if we wish to know what variables should be selected, the situation changes. We argue that a key assumption for model selection consistency for the ℓ_1-penalty (Lasso) does not hold while a SCAD-type penalty does.

Zhao and Yu (2006) studied a concept called sign consistency defined by $P(\mathrm{sgn}(\hat{\boldsymbol{\beta}}) - \mathrm{sgn}(\boldsymbol{\beta}_0)) \to 1$, which is stronger than model selection consistency. Under a deterministic covariate assumption, they show that the *weak irrepresentable condition*

$$\|\boldsymbol{H}_{BAT}\boldsymbol{H}_{AAT}^{-1}\mathrm{sgn}(\boldsymbol{\beta}_{0A})\|_\infty < 1$$

is necessary for the sign consistency of Lasso. To establish the model selection consistency of Lasso, we usually need a stronger condition

$$\|\boldsymbol{H}_{BAT}\boldsymbol{H}_{AAT}^{-1}\|_\infty \leq C \ \text{ for some } \ C \in (0,1),$$

which was supposed by Fan and Lv (2011). It seems difficult to prove model selection consistency for the Lasso without this condition; however, the condition may be easily violated. Let $\boldsymbol{x}_i$, $i \in B$, be a column vector of $\boldsymbol{X}_B$. Then, the left-hand side of the bound is

$$\|\boldsymbol{H}_{BAT}\boldsymbol{H}_{AAT}^{-1}\|_\infty = \max_{i\in B} \|(\boldsymbol{X}_A^\top \boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^\top \boldsymbol{x}_i\|_1 =: \max_{i\in B} \|\hat{\boldsymbol{\pi}}_i\|_1,$$

where $\hat{\boldsymbol{\pi}}_i \in \mathbb{R}^q$ is regarded as the OLS estimator of regression of an irrelevant variable $\boldsymbol{x}_i$ on important variables $\boldsymbol{X}_A$. Due to stationarity, this is $O_p(q)$ provided that the regularity conditions for an asymptotic theory are satisfied. Even when q is finite, it is unrealistic for this value to be strictly bounded by one since macroeconomic data have cross-sectional dependence in general. When lagged variables are included in $\boldsymbol{X}$, the condition becomes more tight because A and B may share the same variable. Violation of the condition would lead to a collapse of economic interpretation of estimated coefficients with the Lasso.

A.3 Lemmas for Theorems 1

The following lemmas were given by Loh and Wainwright (2015, Lemma 4(b) and Lemma 5), and are consequences of Assumption 2. They are used to fill the gap between the ℓ_1-norm and SCAD-type penalties. The proofs are omitted.

Lemma 1 *Under Assumotion 2, any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfies*

$$\lambda\|\boldsymbol{\beta}\|_1 \leq \|p_\lambda(\boldsymbol{\beta})\|_1 + (\mu/2)\|\boldsymbol{\beta}\|_2^2.$$

Lemma 2 *Under Assumotion 2, for any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\xi p_\lambda(\boldsymbol{\beta}_0) - p_\lambda(\boldsymbol{\beta}) > 0$ and $\xi \geq 1$, we have*

$$\xi\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta})\|_1 \leq \xi\lambda\|\boldsymbol{\beta}_A - \boldsymbol{\beta}_{0A}\|_1 - \lambda\|\boldsymbol{\beta}_B - \boldsymbol{\beta}_{0B}\|_1.$$

A.4 Lemmas for Theorem 2

In Lemma 3 below, let $\hat{A} := \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}$, a set of indices corresponding to all nonzero components of $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}}_{\hat{A}}$ denote a subvector of $\hat{\boldsymbol{\beta}}$ formed by its restriction to $\hat{A}$. The other symbols are defined analogously. Let $\circ$ denote the Hadamard product. The sign function $\mathrm{sgn}(\cdot)$ is applied coordinate-wise. Define

$$G_{\hat{A}T}(\hat{\boldsymbol{\beta}}) = -T^{-1}\boldsymbol{X}_{\hat{A}}^{\top}\boldsymbol{y} + T^{-1}\boldsymbol{X}_{\hat{A}}^{\top}\boldsymbol{X}_{\hat{A}}\hat{\boldsymbol{\beta}}_{\hat{A}},$$
$$G_{\hat{B}T}(\hat{\boldsymbol{\beta}}) = -T^{-1}\boldsymbol{X}_{\hat{B}}^{\top}\boldsymbol{y} + T^{-1}\boldsymbol{X}_{\hat{B}}^{\top}\boldsymbol{X}_{\hat{A}}\hat{\boldsymbol{\beta}}_{\hat{A}}.$$

Define the *local concavity* at $\boldsymbol{b} \in \mathbb{R}^r$ with $\|\boldsymbol{b}\|_0 = r$ as $\kappa_\lambda(\boldsymbol{b}) = \max_{1 \leq j \leq r} -p''_\lambda(|b_j|)$.

Lemma 3 *Suppose Assumption 2 holds. Then $\hat{\boldsymbol{\beta}}$ is a strict local minimizer of $Q_T(\boldsymbol{\beta})$ in (2) if*

$$G_{\hat{A}T}(\hat{\boldsymbol{\beta}}) + p'_\lambda(\hat{\boldsymbol{\beta}}_{\hat{A}}) \circ \mathrm{sgn}(\hat{\boldsymbol{\beta}}_{\hat{A}}) = 0, \tag{6}$$
$$\|G_{\hat{B}T}(\hat{\boldsymbol{\beta}})\|_\infty < p'_\lambda(0+), \tag{7}$$
$$\Lambda_{\min}(\boldsymbol{H}_{\hat{A}\hat{A}T}) > \kappa_\lambda(\hat{\boldsymbol{\beta}}_{\hat{A}}). \tag{8}$$

Conversely, any local minimizer of $Q_T(\boldsymbol{\beta})$ must satisfy (6), (7), and (8) with strict inequalities replaced by nonstrict ones.

The proof was given by Lv and Fan (2009, Theorem 1). Consider the case where $\hat{\boldsymbol{\beta}}_{\hat{A}} \in \mathcal{N}_0$. Under Assumption 8, it holds that $\sup_{\boldsymbol{\beta}_A \in \mathcal{N}_0} \kappa_\lambda(\boldsymbol{\beta}_A) = 0$ for sufficiently large T. Thus, condition (8) is satisfied as long as $\Lambda_{\min}(\boldsymbol{H}_{\hat{A}\hat{A}T})$ is bounded away from zero.

A.5 Proofs of Theorems 1 and 2

Proof of Theorem 1 Because $\hat{\boldsymbol{\beta}}$ minimizes $Q_T(\boldsymbol{\beta})$, we have

$$(2T)^{-1}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 + \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 \leq (2T)^{-1}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0\|_2^2 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1.$$

By model (1) and Holder's inequality, this can be rewritten and bounded as

$$(2T)^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq T^{-1}\boldsymbol{u}^{\top}\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1$$

$$\leq \|T^{-1}\boldsymbol{X}^\top \boldsymbol{u}\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1. \tag{9}$$

In what follows, we have only to work on event $\mathcal{E}_1$ defined in Assumption 3. On the event, we have $\|T^{-1}\boldsymbol{X}^\top \boldsymbol{u}\|_\infty \leq \lambda/2$, so that (9) becomes

$$(2T)^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq 2^{-1}\lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1. \tag{10}$$

By Lemma 1, the first term in the upper bound of (10) is further bounded by

$$\begin{aligned}\lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 &\leq \|p_\lambda(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_1 + (\mu/2)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \\ &\leq \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 + (\mu/2)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2,\end{aligned} \tag{11}$$

where the last inequality follows from the subadditivity implied by the concavity of the penalty function. On the other hand, since $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_0 \leq \|\hat{\boldsymbol{\beta}}\|_0 + \|\boldsymbol{\beta}_0\|_0 \leq m$ holds on the assumed parameter space due to $\|\boldsymbol{\beta}_0\|_0 = s$, Assumption 4 yields the lower bound of (10); that is, we have on $\mathcal{E}_2$ defined in Assumption 4

$$T^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \geq \gamma\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2. \tag{12}$$

Therefore, combining (10) with (11) and (12) gives

$$(\gamma - \mu/2)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq 3\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1. \tag{13}$$

In particular, (13) implies $3\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 \geq 0$, so that we can apply Lemma 2 to the right-hand side of (13) to obtain

$$(\gamma - \mu/2)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq 3\lambda\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 - \lambda\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{0B}\|_1. \tag{14}$$

Ignoring the last term and the Cauchy-Schwarz inequality lead to

$$(\gamma - \mu/2)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq 3\lambda\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 \leq 3s^{1/2}\lambda\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_2 \leq 3s^{1/2}\lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2,$$

which concludes the error bound in the ℓ_2-norm

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq \frac{6s^{1/2}\lambda}{2\gamma - \mu}. \tag{15}$$

Using (15), we can obtain the error bound in the ℓ_1-norm as well. Since (14) also implies that $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{0B}\|_1 \leq 3\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1$, we have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 &= \|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 + \|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_{0B}\|_1 \\ &\leq 4\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_1 \leq 4s^{1/2}\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_2 \leq 4s^{1/2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq \frac{24s\lambda}{2\gamma - \mu}. \end{aligned} \tag{16}$$

Finally, we derive the prediction error bound from (16). The Mean value theorem, Assumption 2, and the triangle inequality give

$$\begin{aligned}\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 &= \sum_{j=1}^{p} \left(|p_\lambda(\beta_{0j})| - |p_\lambda(\hat{\beta}_j)|\right) = \sum_{j=1}^{p} p'_\lambda(b_j)\left(|\beta_{0j}| - |\hat{\beta}_j|\right) \\ &\leq p'_\lambda(0+)\sum_{j=1}^{p}\left||\beta_{0j}| - |\hat{\beta}_j|\right| \leq \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1,\end{aligned}$$

where b_j is a point between $|\beta_{0j}|$ and $|\hat{\beta}_j|$. Hence, using (10), we obtain

$$T^{-1}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq 3\lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq \frac{72s\lambda^2}{2\gamma - \mu}. \tag{17}$$

Results (15)–(17) hold with probability at least $1 - O(p^{-c_1}) - O(\exp(-c_2 T)$ by Assumptions 3 and 4. □

Proof of Theorem 2 First, we show results (a) and (b) through the following steps.

Step 1. We consider $Q_T(\boldsymbol{\beta})$ in the correctly constrained space $\{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_B = \mathbf{0} \in \mathbb{R}^{p-s}\}$, which is the s-dimensional subspace $\{\boldsymbol{\beta}_A \in \mathbb{R}^s\}$. The corresponding objective function is given by

$$Q_T(\boldsymbol{\beta}_A, \mathbf{0}) = (2T)^{-1}\|\boldsymbol{y} - \boldsymbol{X}_A\boldsymbol{\beta}_A\|_2^2 + \|p_\lambda(\boldsymbol{\beta}_A)\|_1. \tag{18}$$

We now show the existence of a strict local minimizer $\hat{\boldsymbol{\beta}}_{0A}$ of $Q_T(\boldsymbol{\beta}_A, \mathbf{0})$ such that $\|\hat{\boldsymbol{\beta}}_{0A} - \boldsymbol{\beta}_{0A}\| = O_p((s/T)^{1/2})$. To this end, it is sufficient to prove that, for a large constant $C > 0$, the event

$$\mathcal{E}_Q = \left\{\inf_{\|\boldsymbol{v}\|_2 = C} Q_T(\boldsymbol{\beta}_{0A} + \boldsymbol{v}(s/T)^{1/2}, \mathbf{0}) > Q_T(\boldsymbol{\beta}_{0A}, \mathbf{0})\right\} \tag{19}$$

occurs with probability tending to one. This implies that, with high probability, there is a local minimizer $\hat{\boldsymbol{\beta}}_{0A}$ of $Q_T(\boldsymbol{\beta}_A, \mathbf{0})$ in the ball $\mathcal{N}_C \equiv \{\boldsymbol{\beta}_A \in \mathbb{R}^s : \|\boldsymbol{\beta}_A - \boldsymbol{\beta}_{0A}\|_2 \leq C(s/T)^{1/2}\}$.

By the definition of the objective function, we have

$$\begin{aligned} R_T(\boldsymbol{v}) &:= Q_T(\boldsymbol{\beta}_{0A} + \boldsymbol{v}(s/T)^{1/2}, \mathbf{0}) - Q_T(\boldsymbol{\beta}_{0A}, \mathbf{0}) \\ &= (s/T)^{1/2}\boldsymbol{v}^\top \boldsymbol{G}_{0AT} + (s/T)\boldsymbol{v}^\top \boldsymbol{H}_{AAT}\boldsymbol{v} \qquad (20) \\ &\quad + \|p_\lambda(\boldsymbol{\beta}_{0A} + \boldsymbol{v}(s/T)^{1/2})\|_1 - \|p_\lambda(\boldsymbol{\beta}_{0A})\|_1. \qquad (21) \end{aligned}$$

First, we evaluate the two terms in (21). The Mean value theorem gives

$$\begin{aligned} \|p_\lambda(\boldsymbol{\beta}_{0A} + \boldsymbol{v}(s/T)^{1/2})\|_1 - \|p_\lambda(\boldsymbol{\beta}_{0A})\|_1 &= \sum_{j\in A} p'_\lambda(|\beta^\star_{0j}|)(|\beta_{0j} + v_j(s/T)^{1/2}| - |\beta_{0j}|) \\ &\leq p'_\lambda(d)(s/T)^{1/2}\|\boldsymbol{v}\|_1, \qquad (22) \end{aligned}$$

where $|\beta^\star_{0j}|$ lies between $|\beta_{0j}|$ and $|\beta_{0j} + v_j(s/T)^{1/2}|$, and the last inequality follows from the monotonicity of $p'_\lambda(\cdot)$, $\min_{j\in A}|\beta^\star_{0j}| \geq d$, and the triangle inequality. Eventually, the last term is zero by Assumption 8. Next, we consider (20). Since martingale difference sequences are serially uncorrelated, Assumption 9 entails that

$$\begin{aligned} \mathrm{E}\|\boldsymbol{G}_{0AT}\|_2^2 &= T^{-2}\mathrm{E}[\boldsymbol{u}^\top \boldsymbol{X}_A \boldsymbol{X}_A^\top \boldsymbol{u}] = T^{-2}\sum_{j\in A}\mathrm{E}[\boldsymbol{u}^\top \boldsymbol{x}_j \boldsymbol{x}_j^\top \boldsymbol{u}] \\ &= T^{-2}\sum_{j\in A}\mathrm{E}\left[\left(\sum_{t=1}^T x_{tj}u_t\right)^2\right] = T^{-2}\sum_{j\in A}\sum_{t=1}^T \mathrm{E}\left[\left(x_{tj}u_t\right)^2\right] = O(s/T). \end{aligned}$$

This together with the Markov inequality implies that $\|\boldsymbol{G}_{0AT}\|_2$ is $O_p((s/T)^{1/2})$. Therefore, the Cauchy-Schwarz inequality yields

$$(s/T)^{1/2}|\boldsymbol{v}^\top \boldsymbol{G}_{0AT}| \leq (s/T)^{1/2}\|\boldsymbol{v}\|_2\|\boldsymbol{G}_{0AT}\|_2 = O_p(s/T)\|\boldsymbol{v}\|_2.$$

Whereas, by Assumption 10, we get

$$(s/T)\boldsymbol{v}^\top \boldsymbol{H}_{AAT}\boldsymbol{v} \geq (s/T)\Lambda_{\min}(\boldsymbol{H}_{AAT})\|\boldsymbol{v}\|_2^2 \geq (s/T)c_H\|\boldsymbol{v}\|_2^2. \qquad (23)$$

Because (23) dominates the other terms of $R_T(\boldsymbol{v})$ when a large value of $\|\boldsymbol{v}\|_2$ is taken, $\inf_{\|\boldsymbol{v}\|_2=C} R_T(\boldsymbol{v})$ tends to positivity as T grows large. Thus, with probability approaching one, (19) holds, and $\|\hat{\boldsymbol{\beta}}_{0A} - \boldsymbol{\beta}_{0A}\|_2 \leq C(s/T)^{1/2}$.

Step 2. To complete the proof of (a) and (b), it remains to show that $\hat{\boldsymbol{\beta}}_0 := (\hat{\boldsymbol{\beta}}_{0A}, \mathbf{0})$ is indeed a strict local maximizer of $Q_T(\boldsymbol{\beta})$ in $\mathbb{R}^p$. From Lemma 3, it suffices to check conditions (6), (7), and (8) with setting $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0$, but condition (6) is satisfied by the proof of Theorem 1 in Fan and Lv (2011).

We then check Condition (8). Define $\mathcal{N}_0 := \{\boldsymbol{\beta}_A \in \mathbb{R}^s : \|\boldsymbol{\beta}_A - \boldsymbol{\beta}_{0A}\|_\infty \leq d\}$, where we recall $d = \min_{j\in A} |\beta_{0,j}|/2$. By Assumption 8, we have $d/(s/T)^{1/2} \to \infty$, so that, for sufficiently large T, $\hat{\boldsymbol{\beta}}_A \in \mathcal{N}_C$ implies $\hat{\boldsymbol{\beta}}_A \in \mathcal{N}_0$. Thus the condition is eventually satisfied by Assumptions 10 and 11 along with the comment after Lemma 3.

To verify (7), we first see that $(s/T)^{1/2}/\lambda = o(1)$ by Assumption 8. Thus, Assumptions 7 and 12 establish

$$\begin{aligned}
\|G_{BT}(\hat{\boldsymbol{\beta}})\|_\infty &= \|\boldsymbol{H}_{BAT}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) + \boldsymbol{G}_{0BT}\|_\infty \leq \|\boldsymbol{H}_{BAT}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A})\|_\infty + \|\boldsymbol{G}_{0BT}\|_\infty \\
&\leq \|\boldsymbol{H}_{BAT}\|_{2,\infty}\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_2 + \lambda/2 \\
&= O_p(1)C(s/T)^{1/2} + \lambda/2 = \left\{o_p(1) + 1\right\}\lambda/2.
\end{aligned}$$

Since $p'_\lambda(0+) = \lambda$ in Assumption 2, condition (7) holds for a sufficiently large T. This completes the proof of (a) and (b).

Finally, we prove (c). Clearly we only need to show the asymptotic normality of $\hat{\boldsymbol{\beta}}_A$. Assumption 11 ensures that $\boldsymbol{I}_{0AA}$ is positive definite, and hence, $\boldsymbol{I}_{0AA}^{-1/2}$ is well-defined. On the event $\mathcal{E}_Q$ in (19), it has been shown that $\hat{\boldsymbol{\beta}}_A \in \mathcal{N}_C$ is a strict local minimizer of $Q_T(\boldsymbol{\beta}_A, \mathbf{0})$ and $\partial Q_T(\hat{\boldsymbol{\beta}}_A, \mathbf{0})/\partial\boldsymbol{\beta}_A = \mathbf{0}$. We thus obtain, for any vector $\boldsymbol{b} \in \mathbb{R}^s$ such that $\|\boldsymbol{b}\|_2 = 1$,

$$-T^{1/2}\boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2}\boldsymbol{H}_{AAT}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) = T^{1/2}\boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2}\boldsymbol{G}_{0AT} + T^{1/2}\boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2} p'_\lambda(\hat{\boldsymbol{\beta}}_A) \circ \operatorname{sgn}(\hat{\boldsymbol{\beta}}_A). \tag{24}$$

Recall that $T^{1/2}\boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2}\boldsymbol{G}_{0AT} = \sum_{t=1}^T \xi_{Tt}$ and ξ_{Tt} is a martingale difference array. We show the

asymptotic normality of this part. It is not hard to say that

$$\sum_{t=1}^{T} \mathrm{Var}(\xi_{Tt}) = \boldsymbol{b}^\top \boldsymbol{I}_{0AA}^{-1/2} \boldsymbol{I}_{0AA} \boldsymbol{I}_{0AA}^{-1/2} \boldsymbol{b} = 1.$$

Assumption 13 implies uniform integrability of ξ_t^2. Hence, by Theorems 24.3 and 24.4 of Davidson (1994, Ch. 24), we obtain $\sum_{t=1}^{T} \xi_{Tt} \to_d N(0,1)$. Because the last term of (24) is $o_p(1)$ by the argument above, the result follows from the Slutsky lemma and Assumption 10. □

A.6 Lemmas for Proposition 1

Recall that $c_{xu} = \limsup_T \max_{t,i}\{R_{tT}^{(X)}\Sigma_{pi}, R_{tT}^{(u)}\Sigma_{pi}\} < \infty$.

Lemma 4 *Under Assumption 6, we have for any i and $\alpha > 0$,*

$$P\left(\max_t |x_{ti}u_t| > \alpha\right) \leq 4T \exp\left\{-\alpha/(2c_{xu})\right\}.$$

Proof We see that

$$P\left(|x_{ti}u_t| > \alpha\right) \leq P\left(|x_{ti}| > \alpha^{1/2}\right) + P\left(|u_t| > \alpha^{1/2}\right).$$

We consider the first term. By the construction of x_{ti} with suppressing the superscript, we have

$$P\left(\left|\sum_{s=1}^{T}\sum_{k=1}^{p} r_{ts}\sigma_{ki}z_{sk}\right| > \alpha^{1/2}\right) = P\left((R_{tT}\Sigma_{pi})^{-1/2}\left|\sum_{s=1}^{T}\sum_{k=1}^{p} r_{ts}\sigma_{ki}z_{sk}\right| > (\alpha/R_{tT}\Sigma_{pi})^{1/2}\right)$$
$$\leq 2\exp\left\{-\alpha/(2R_{tT}\Sigma_{pi})\right\} \leq 2\exp\left\{-\alpha/(2c_{xu})\right\},$$

where the first inequality holds since $(R_{tT}\Sigma_{pi})^{-1/2}\sum_{s=1}^{T}\sum_{k=1}^{p} r_{ts}\sigma_{ki}z_{sk}$ is a standard normal random variable and the last inequality follows from Assumption 6. It is clear that we obtain the same result for u_t. Therefore, by the union bound, we have

$$P\left(\max_t |x_{ti}u_t| > \alpha\right) \leq 4T \exp\left\{-\alpha/(2c_{xu})\right\},$$

which yields the desired inequality. □

Lemma 5 *Let $\lambda = c_0 \log(pT)(\log p/T)^{1/2}$ for any positive constants c_0. Let m_0 be an arbitrary positive constant. Under Assumption 6, we have for any i,*

$$P\left(T^{-1}|\boldsymbol{x}_i^\top \boldsymbol{u}| \geq \lambda/2 \,\middle|\, \max_t |x_{ti}u_t| \leq m_0 \log(pT)\right) \leq 2p^{-c_0^2/(8m_0^2)}.$$

Proof Because $(x_{ti}u_t, \mathcal{F}_t)$ is a martingale difference sequence with respect to $\mathcal{F}_t = \{u_{t-j}, x_{t-j+1} : j = 0, 1, \dots\}$ for each i, Azuma-Hoeffding's inequality yields

$$\begin{aligned} P\left(T^{-1}|\boldsymbol{x}_i^\top \boldsymbol{u}| \geq \lambda/2 \,\middle|\, \max_t |x_{ti}u_t| \leq \alpha\right) &= P\left(\left|\sum_{t=1}^T x_{ti}u_t\right| \geq T\lambda/2 \,\middle|\, \max_t |x_{ti}u_t| \leq \alpha\right) \\ &\leq 2\exp\left[-\frac{(T\lambda/2)^2}{2T\alpha^2}\right] \end{aligned}$$

for any $\alpha > 0$ for each T and p. Plugging λ and $\alpha = m_0 \log(pT)$ into the upper bound, we have

$$2\exp\left[-\frac{Tc_0^2(\log pT)^2 \log p}{8Tm_0^2(\log pT)^2}\right] = 2\exp\left[-\frac{c_0^2 \log p}{8m_0^2}\right] = 2p^{-c_0^2/(8m_0^2)},$$

giving the result. □

A.7 Proofs of Propositions 1 and 2

Proof of Proposition 1 By the union bound and the property of the conditional probability, we have for any $\alpha > 0$,

$$\begin{aligned} P(\mathcal{E}_1^c) &= P(T^{-1}\max_i |\boldsymbol{x}_i^\top \boldsymbol{u}| \geq \lambda/2) \leq \sum_{i=1}^p P(T^{-1}|\boldsymbol{x}_i^\top \boldsymbol{u}| \geq \lambda/2) \\ &\leq \sum_{i=1}^p P(T^{-1}|\boldsymbol{x}_i^\top \boldsymbol{u}| \geq \lambda/2 \,|\, \max_t |x_{ti}u_t| \leq \alpha) + \sum_{i=1}^p P(\max_t |x_{ti}u_t| > \alpha). \end{aligned}$$

Let $\alpha = m_0 \log(pT)$ be the same as in the proof of Lemma 5. From Lemmas 4 and 5, this is bounded as

$$\begin{aligned} P(\mathcal{E}_1^c) &\leq 2p^{-\{c_0^2/(8m_0^2)-1\}} + 4pT\exp\{-m_0 \log(pT)/(2c_{xu})\} \\ &\leq 2p^{-c_0^2/(8m_0^2)+1} + 4(pT)^{-m_0/(2c_{xu})+1}. \end{aligned}$$

Since m_0 and c_0 are arbitrary, putting $m_0 = c_0/4$ and $c_0 \geq 16c_{xu}$ reduces to

$$P(\mathcal{E}_1^c) \leq 2p^{-1} + 4(pT)^{-c_0/(8c_{xu})+1} \leq 2p^{-1} + 4(pT)^{-1} \leq 6p^{-1},$$

giving the result. □

Proof of Proposition 2 We have

$$m \leq \phi T^{1-\delta} < T, \tag{25}$$

where the last strict inequality holds for large T. Note that each row of $\boldsymbol{W} \equiv \boldsymbol{Z}\boldsymbol{\Sigma}_X^{1/2}$ is viewed as an p-dimensional random vector independently sampled from $N(0, \boldsymbol{\Sigma}_X)$. Let $\tilde{\boldsymbol{d}} = \boldsymbol{d}_{\mathrm{supp}(d)}$, $\tilde{\boldsymbol{X}} = \boldsymbol{X}_{\mathrm{supp}(d)}$ and $\tilde{\boldsymbol{W}} = \boldsymbol{W}_{\mathrm{supp}(d)}$. For any $\mathrm{supp}(\boldsymbol{d}) \subset \{1, \ldots, p\}$ satisfying (25) and $\|\boldsymbol{d}\|_0 \leq m$, we see that

$$\begin{aligned}
T^{-1}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{d}}\|_2^2/\|\tilde{\boldsymbol{d}}\|_2^2 &= T^{-1}\left(\frac{\tilde{\boldsymbol{d}}^\top\tilde{\boldsymbol{W}}^\top\boldsymbol{R}_X\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}}{\tilde{\boldsymbol{d}}^\top\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}}\right)\left(\frac{\tilde{\boldsymbol{d}}^\top\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}}{\tilde{\boldsymbol{d}}^\top\tilde{\boldsymbol{d}}}\right) \\
&\geq T^{-1}\min_{\boldsymbol{h}\in\mathbb{R}^T}\left(\frac{\boldsymbol{h}^\top\boldsymbol{R}_X\boldsymbol{h}}{\boldsymbol{h}^\top\boldsymbol{h}}\right)\min_{\tilde{\boldsymbol{d}}\in\mathbb{R}^m}\left(\frac{\tilde{\boldsymbol{d}}^\top\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}}{\tilde{\boldsymbol{d}}^\top\tilde{\boldsymbol{d}}}\right) \\
&\geq c_R\min_{\tilde{\boldsymbol{d}}\in\mathbb{R}^m} T^{-1}\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}\|_2^2/\|\tilde{\boldsymbol{d}}\|_2^2,
\end{aligned} \tag{26}$$

where the last inequalities hold by Assumption 6. We denote by $\tilde{\boldsymbol{w}}_t$ and $\tilde{c}$ the tth row of $\tilde{\boldsymbol{W}}$ and the minimum eigenvalue of the covariance matrix of $\tilde{\boldsymbol{w}}_t$, respectively. Inequality (26) with the fact that $\tilde{c} \geq c_\Sigma$ leads to

$$\begin{aligned}
P\left(\min_{\tilde{\boldsymbol{d}}\in\mathbb{R}^m} T^{-1}\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{d}}\|_2^2/\|\tilde{\boldsymbol{d}}\|_2^2 \leq c_\Sigma c_R/9\right) &\leq P\left(c_R\min_{\tilde{\boldsymbol{d}}\in\mathbb{R}^m} T^{-1}\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}\|_2^2/\|\tilde{\boldsymbol{d}}\|_2^2 \leq c_\Sigma c_R/9\right) \\
&\leq P\left(\min_{\tilde{\boldsymbol{d}}\in\mathbb{R}^m} T^{-1}\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}\|_2^2/\|\tilde{\boldsymbol{d}}\|_2^2 \leq \tilde{c}/9\right).
\end{aligned} \tag{27}$$

An application of Lemma 9 in Wainwright (2009) gives

$$P\left(\min_{\tilde{\boldsymbol{d}}\in\mathbb{R}^m} T^{-1}\|\tilde{\boldsymbol{W}}\tilde{\boldsymbol{d}}\|_2^2/\|\tilde{\boldsymbol{d}}\|_2^2 \leq \tilde{c}/9\right) \leq 2\exp(-T/2). \tag{28}$$

Finally, we extend the result uniformly in terms of the choice of supp($\boldsymbol{d}$). We see that $\binom{p}{m} \leq p^m \leq \exp(\phi^2 T)$ holds for large T by Stirling's approximation and (25) with Assumption 5. Therefore, taking the union bound with combining (27) and (28) gives

$$P\left(\min_{\boldsymbol{d}\in\mathbb{R}^p,\, \|\boldsymbol{d}\|_0\leq m} T^{-1}\|\boldsymbol{X}\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_2^2 \leq c_G c_R/9\right) \leq 2\exp(\phi^2 T - T/2),$$

which goes to zero since $\phi^2 < 1/2$ by Assumption 5. Consequently, if we choose $\gamma = c_G c_R/9$ and $c_2 = 1/2 - \phi^2$, we achieve the result. □

A.8 Collinearity

We explore how collinearity between $\boldsymbol{X}_B$ and $\boldsymbol{X}_A$ affects the oracle property obtained by Theorem 2. Assumption 12 controls how much collinearity is allowed. Recall that $\boldsymbol{H}_{BAT} = T^{-1}\boldsymbol{X}_B^\top \boldsymbol{X}_A$ for $\boldsymbol{X}_A \in \mathbb{R}^{T\times s}$ and $\boldsymbol{X}_B \in \mathbb{R}^{T\times(p-s)}$. We are interested in the behavior of

$$\|\boldsymbol{H}_{BAT}\|_{2,\infty} \equiv \max_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{H}_{BAT}\boldsymbol{v}\|_\infty = \max_{b\in B} \max_{\|\boldsymbol{v}\|_2=1} \left|T^{-1}\boldsymbol{x}_b^\top \boldsymbol{X}_A \boldsymbol{v}\right|,$$

where we write $\boldsymbol{X}_A\boldsymbol{v} = \sum_{a\in A} v_a \boldsymbol{x}_a$. This value is expected to become unbounded (and hence Assumption 12 is violated) under strong collinearity.

To obtain understandable results, we make the following simplified assumptions: the regressors are deterministic, and for any $b \in B$ and $a \in A$, $T^{-1}\boldsymbol{x}_b^\top \boldsymbol{x}_a \to \rho_{ba} \geq 0$. Moreover, we assume either of the two conditions:

1. $\max_{b\in B} \rho_{ba} \geq c > 0$ for all $a \in A$,
2. $\max_{b\in B} \rho_{ba} \leq c a^{-q/2}$ for some $q > 1$.

Condition 1 describes a highly correlated case. The correlation between $\boldsymbol{x}_b$ and $\boldsymbol{x}_a$ always exists even if s increases. On the other hand, condition 2 models weaker correlations than condition 1 does. Specifically, most of the correlations become small as q becomes large, meaning that the effect of collinearity is limited in this case. In fact, it is not difficult to see that $\|\boldsymbol{H}_{BAT}\|_{2,\infty}$ diverges at least as fast as $s^{1/2}$ under condition 1 while $\|\boldsymbol{H}_{BAT}\|_{2,\infty}$ is uniformly

bounded under condition 2: First, we suppose condition 1 and let $\bar{\boldsymbol{v}} = (s^{-1/2}, \ldots, s^{-1/2})^\top$. We then observe that

$$\max_{b\in B} \max_{\|\boldsymbol{v}\|_2=1} \left|T^{-1}\boldsymbol{x}_b^\top \boldsymbol{X}_A \boldsymbol{v}\right| \geq \max_{b\in B} \left|T^{-1}\boldsymbol{x}_b^\top \boldsymbol{X}_A \bar{\boldsymbol{v}}\right| = \max_{b\in B} \left| s^{-1/2} \sum_{a\in A} T^{-1}\boldsymbol{x}_b^\top \boldsymbol{x}_a \right|.$$

By condition 1, the last term is bounded from below by

$$s^{-1/2} \max_{b\in B} \left| \sum_{a\in A} (\rho_{ba} + o(1)) \right| \geq s^{1/2}(c - o(1)),$$

which goes to infinity as $s \to \infty$. Next, we suppose condition 2. By the Cauchy-Schwarz inequality, we observe that

$$\begin{aligned}
\max_{b\in B} \max_{\|\boldsymbol{v}\|_2=1} \left|T^{-1}\boldsymbol{x}_b^\top \boldsymbol{X}_A \boldsymbol{v}\right| &= \max_{b\in B} \max_{\|\boldsymbol{v}\|_2=1} \left| \sum_{a\in A} v_a T^{-1}\boldsymbol{x}_b^\top \boldsymbol{x}_a \right| \\
&= \max_{b\in B} \max_{\|\boldsymbol{v}\|_2=1} \left| \sum_{a\in A} v_a \left(\rho_{ba} + o(1)\right) \right| \\
&\leq \max_{b\in B} \left(\sum_{a\in A} \rho_{ba}^2 \left(1 + o(1)\right) \right)^{1/2} \leq c \left(\sum_{a\in A} a^{-q} \left(1 + o(1)\right) \right)^{1/2}.
\end{aligned}$$

The last term converges since $q > 1$ under condition 2.

The following simulation shows that the strong collinearity (condition 1) affects the oracle property. Table 7 shows the *relative* finite sample success rates detecting non-zero (SC-A) coefficients and zero coefficients (SC-B) that are defined as $SC\text{-}A = P\left(\text{sgn}(\hat{\boldsymbol{\beta}}_A) = \text{sgn}(\boldsymbol{\beta}_{0A})\right)$ and SC-$B = P\left(\text{sgn}(\hat{\boldsymbol{\beta}}_B) = \text{sgn}(\boldsymbol{\beta}_{0B})\right)$ respectively, and (average) mean squared error for estimates of non-zero coefficients (MSE$(\hat{\boldsymbol{\beta}}_A)$) under Condition 1 compared to that of Condition 2 when $T = 300, 500, 1000$ and $c = 0.5, 0.98$ with $q = 4$, $p = 1.5\exp(T^{0.31})$ and $s = 20T^{0.3}$. Then, the finite sample properties of estimators under Condition 1 are equivalent to those of Condition 2 if the values in the Table are 1. We can confirm facts from Table 7 that (i) the success rates are relatively low under Condition 1 irrespective of the degree of collinearity (c) and (ii) the MSE of the Condition 1 is expected to be much worse than the that of Condition 2 asymptotically especially when the degree of collinearity is high. These facts are consistent to the theoretical results because the Condition 1 violates Assumption 12 so that the oracle property no longer holds under Condition 1.

Table 7: Relative SC-A, SC-B and MSE (cond.1/cond.2)

	$c = 0.5$			$c = 0.98$		
	$SC-A$	$SC-B$	MSE	$SC-A$	$SC-B$	MSE
$T = 300$	0.89	0.98	1.10	0.96	1.01	0.99
$T = 500$	0.88	0.99	1.19	0.96	1.00	0.98
$T = 1000$	1.00	1.00	1.25	0.95	1.00	3.49

A.9 Related works

Wang et al. (2007) investigated the asymptotic properties of the Lasso and modified Lasso (Lasso*) for the linear regression with the autoregressive error model. They derived the model selection consistency, and showed the Lasso* can be the oracle estimator. Nardi and Rinaldo (2011) considered the estimation and variable (lag) selection of autoregressive models via the Lasso. They mainly focused on the lag selection of the AR parameters. Lasso-type estimation of VAR models has been studied by several authors, including Song and Bickel (2011), Nicolson et al. (2015), Basu and Michailidis (2015), and Kock and Callot (2015). Theoretically, the latter two papers have significant contribution to the high-dimensional time series literature, but their settings are different from ours. The results obtained here are new and much complement their works. Basu and Michailidis (2015) investigated estimation of general high-dimensional time dependent models via spectral densities of covariates and errors, and derived the non-asymptotic error bound. Kock and Callot (2015) derived the non-asymptotic error bound for a high-dimensional VAR model.